ISD Institute for Strategic Dialogue

gefördert durch:
Auswärtiges Amt

News Feed, Reels & »Für Dich«: Wie algorithmische Ranking-Praktiken unser Online-Umfeld beeinflussen und schützen könnten

Sara Bundtzen

Über das Digital Policy Lab

Als zwischenstaatliche Arbeitsgruppe engagiert sich das Digital Policy Lab (DPL) dafür, politische Lösungen zur Verhinderung und Bekämpfung der Verbreitung von Desinformation, Hassrede sowie extremistischen und terroristischen Inhalten im Internet aufzuzeigen. Die Arbeitsgruppe besteht aus Vertreter:innen der zuständigen Ministerien und Aufsichtsbehörden ausgewählter liberal-demokratischer Länder. Die Arbeit des DPL zielt darauf ab, den regierungsübergreifenden Dialog zu fördern, politischen Entscheidungsträger:innen und Aufsichtsbehörden Zugang zu einschlägigem Fachwissen und Forschungsergebnissen zu verschaffen sowie eine internationale Arbeitsgemeinschaft zur Bewältigung der wichtigsten digitalpolitischen Herausforderungen aufzubauen. Wir danken dem Auswärtigen Amt für die Unterstützung des Projekts.

Über diesen Bericht

Im Rahmen des Digital Policy Lab organisierte das ISD im Juli 2022 zwei Arbeitsgruppentreffen zum Thema algorithmischer Ranking-Systeme. Die Teilnehmer:innen vertraten Ministerien, Ämter und Aufsichtsbehörden aus Australien, Kanada, Deutschland, Irland, den Niederlanden, Neuseeland, Großbritannien und den USA. Zu den Teilnehmer:innen gehörten neben Vertreter:innen aus Wissenschaft und Zivilgesellschaft auch ehemalige Mitarbeiter:innen von Plattformen. Die Arbeitsgruppe widmete sich den gesellschaftlichen Auswirkungen algorithmischer Ranking-Praktiken der Online-Plattformen sowie möglichen Ansätzen für mehr Verantwortung seitens der Unternehmen. Dieser Bericht baut auf den Diskussionen auf und vertieft sie. Die hierin geäußerten Auffassungen spiegeln nicht zwangsläufig die Ansichten der Teilnehmer:innen oder der an diesem Projekt beteiligten Staaten wider.

Autorin

Sara Bundtzen ist Analystin bei ISD Germany. Sie erforscht die Verbreitung von Desinformation und Informationsmanipulation durch staatliche und nichtstaatliche Akteur:innen im deutschen und englischen Sprachraum. Im Rahmen des Digital Policy Lab (DPL) unterstützt Sara Bundtzen die digitalpolitische Arbeit des ISD und untersucht Vorschläge zur Bekämpfung von Desinformation, Einflusskampagnen, Hassrede und extremistischen Inhalten.

Herausgeberische Verantwortung;
Huberta von Voss, Executive Director ISD Germany

Danksagungen

Unser ausdrücklicher Dank gilt den Teilnehmer:innen der Zivilgesellschaft und der Wissenschaft für ihre wertvollen Beiträge in der Arbeitsgruppe: Sahar Massachi (Integrity Institute), Jenny Brennan (Ada Lovelace Institute), Professor Barak Richman (Duke University), Dr. Anna-Katharina Meßmer (Stiftung Neue Verantwortung), Oliver Marsh (The Data Skills Consultancy) und Marie-Therese Sekwenz (TU Delft). Die Autorin dankt Henry Tuck und Helena Schwertheim für deren Feedback und Änderungsvorschläge.

ISD | Institute for Strategic Dialogue

www.isdgermany.org

Inhaltsverzeichnis

Executive Summary

Die meisten Social-Media-Plattformen bieten heute weit mehr Inhalte an als deren Nutzer:innen tatsächlich aufnehmen können. Angesichts der zunehmenden Menge an Inhalten und Formaten bei gleichbleibender Aufmerksamkeitsspanne sind viele Plattformen von umgekehrt chronologischen Feeds zu algorithmischen Ranking-Systemen übergegangen. Diese sollen den Nutzer:innen nicht mehr die aktuellsten, sondern die vermeintlich für sie »interessantesten« Inhalte anzeigen. Als »interessant« werden dabei in der Regel die Inhalte eingestuft, die den höchsten prognostizierten Wert für ein Unternehmen haben – beispielsweise, weil sie die Nutzungsdauer der Nutzer:innen auf einer Plattform verlängern.

Algorithmische Ranking-Systeme, die auch als Empfehlungs- oder Recommender-Systeme bezeichnet werden, treffen automatische Entscheidungen beispielsweise darüber, welche Inhalte im Feed oder in den Suchergebnissen bevorzugt bzw. herabgestuft werden, mit wem man sich vernetzen und wem oder welchen Seiten man folgen soll. Auf diese Weise prägen sie letztlich die Online-Erfahrung von Milliarden von Nutzer:innen. Aufgrund der Schnelligkeit, Reichweite und Zugänglichkeit der Informationen haben sich Social-Media-Plattformen als fester Bestandteil der täglichen Kommunikation und des Nachrichtenkonsums etabliert. Gleichzeitig finden irreführende, hasserfüllte, verschwörungsideologische oder extremistische Ansichten oft ihren Weg in die öffentliche Debatte noch bevor sachliche oder differenzierte Informationen überhaupt vorliegen. Dies führt zu einer erheblichen Verzerrung des Nachrichtenkonsums und der gesellschaftlichen Debatten.

In diesem Bericht werden der Einsatz algorithmischer Ranking-Praktiken sowie deren Auswirkungen auf Online-Diskurse untersucht. Er befasst sich mit dem sogenannten »Engagement-Problem« sowie anderen algorithmischen Phänomenen und wie diese die Verbreitung schädlicher oder grenzwertiger Inhalte verstärken und dabei Vorurteile und Diskriminierung festigen können. Angesichts einer lückenhaften Wissensbasis mit Blick auf die gesellschaftliche Gesamtwirkung der eingesetzten algorithmischen Ranking-Systeme zeigt der vorliegende Bericht die methodischen und epistemischen Herausforderungen – sowohl innerhalb der Forschungsgemeinschaft als auch in Bezug auf regulatorische Rahmenbedingungen – auf, die mit der Auditierung von Algorithmen durch unabhängige Dritte verbunden sind. Der Bericht beleuchtet mögliche Schritte für mehr Transparenz und den Aufbau gemeinsamer Qualitätsstandards zur Durchführung unabhängiger Audits.

Mit Blick auf mögliche Interventionen seitens der Plattformen sowie der Aufsichtsbehörden werden in diesem Bericht die Vor- und Nachteile neuer Ansätze untersucht. Es werden dabei eine stärkere Selbstbestimmung der Nutzer:innen, der Einsatz von sogenannter Middleware, sowie die Entwicklung »qualitätsorientierter« und »brückenbildender« Ranking-Algorithmen diskutiert. Einige dieser Ansätze zielen dabei auf eine Neubewertung der Anreize und Ziele hinter den Metriken ab, die Unternehmen zum Test und zur Bewertung der Effektivität von Ranking-Algorithmen verwenden. Damit gehen sie über Vorschläge hinaus, die vor allem auf eine individualistische Stärkung der Selbstbestimmung auf Seiten der Nutzer:innen setzen. Berücksichtigt werden bekannte Branchenpraktiken zur Selbstregulierung als auch von liberal-demokratischen Staaten geplante Regulierungsvorhaben.

Glossar

Algorithmen bezeichnen in der Informatik üblicherweise eine endliche Abfolge genau definierter, für den Computer ausführbarer Befehle zur Lösung einer Klasse von Problemen oder zur Durchführung einer Berechnung.[1] Algorithmen umfassen einfache if-then-Anweisungen sowie Abfolgen komplexerer mathematischer Modelle, darunter Algorithmen für maschinelles Lernen, neuronale Netze und Deep-Learning-Algorithmen.

Application Programming Interfaces (APIs) sind Softwareschnittstellen, die eine Kommunikation zwischen zwei Anwendungen ermöglichen. Als solche ermöglichen APIs Forscher:innen den Zugriff auf bestimmte Daten von Online-Plattformen über Datenanfragen. Als zwischengeschaltete Instanz stellen APIs eine zusätzliche Sicherheitsebene bereit, indem sie das Volumen und die Häufigkeit der Anfragen protokollieren, verwalten und überwachen.

Engagement- bzw. Interaktionsraten *(engagement rates)* beschreiben den Grad der Interaktion der Nutzer:innen in Bezug auf einen bestimmten Inhalt. Die dazugehörigen Metriken können Interaktionen wie das »Liken« oder andere Reaktionen auf Inhalte, das Kommentieren oder Teilen von Beiträgen, das Abrufen eines Fotos oder Videos oder das Anklicken von Links erfassen. Im Gegensatz dazu sind Impressionen (*impressions*) die Anzahl der Anzeigen eines Inhalts auf der Benutzeroberfläche (z. B. im Feed), unabhängig davon, ob die entsprechenden Nutzer:innen den Inhalt tatsächlich gesehen (*reach* bzw. Reichweite) oder direkt mit ihm interagiert haben (Engagement).

Inhaltsmoderation (*content moderation*) bezeichnet Governance-Mechanismen zur Durchsetzung von Regeln in Bezug auf die Aktivitäten auf einer Plattform. Bei Inhaltsmoderator:innen handelt es sich entweder um Angestellte des Plattformbetreibers oder um externe Auftragnehmer:innen. Ihre Aufgabe besteht darin, Inhalte, die gegen die Nutzungsbedingungen einer Plattform verstoßen, zu erkennen, zu überprüfen, herabzustufen und/oder zu entfernen. Um mit dem wachsenden Umfang der Inhalte Schritt halten zu können, setzen Plattformen oder Anbieter von Moderationsdiensten zunehmend auf automatisierte Systeme zur Inhaltsmoderation sowie auf automatisierte Tools zur Unterstützung derInhaltsmoderator:innen.

Inhaltsempfehlungen sind von Algorithmen priorisierte Vorschläge, die den Nutzer:innen in Form von Beiträgen, Seiten, Gruppen oder Accounts angezeigt werden. Bekannte Beispiele von algorithmischen Empfehlungssystemen sind der Explore-Feed von Instagram, YouTube Shorts (Kurzvideo-Formate), der Für-Dich-Feed von TikTok oder der News Feed von Facebook.

Maschinelles Lernen (ML) ist ein Teilgebiet künstlicher Intelligenz (KI), die es Softwareanwendungen ermöglicht, Ergebnisse genauer vorherzusagen, ohne dass sie explizit dafür programmiert wurden. ML-Algorithmen verwenden historische Daten als Eingabe, um neue Ausgaben vorherzusagen. Überwachtes ML erfordert, dass Datenwissenschaftler:innen die Algorithmen sowohl mit gekennzeichneten Eingaben als auch mit den gewünschten Ausgaben trainieren.[2]

Schädliche Inhalte und Verhaltensweisen (*harmful content and behaviour*) umfassen ein breites Spektrum von Online-Aktivitäten, die negative Auswirkungen auf den demokratischen und gesellschaftlichen Diskurs haben können. Dazu zählen Inhalte in Form von Hassrede, Anstiftung zur Gewalt gegen eine bestimmte Gruppe, verschwörungsideologische Inhalte, sowie falsche, irreführende oder manipulierte Inhalte. In einigen Fällen kann das Schadensrisiko direkt von dem betreffenden Inhalt ausgehen. In anderen Fällen wird das Schadensrisiko nicht durch den Inhalt als solchen, sondern durch kollektive Verhaltensmuster verursacht. In beiden Fällen können die Risiken durch eine künstliche Verstärkung verschärft werden. Je nach rechtlichem Rahmen können verschiedene Formen von schädlichen Inhalten oder Verhaltensweisen rechtswidrig sein oder nicht. Je nach Plattform können schädliche Inhalte oder Verhaltensweisen in den Geltungsbereich der Community-Richtlinien, Standards oder unternehmenseigenen Regeln fallen oder auch nicht. Einige Unternehmen sprechen von grenzwertigen Inhalten (*borderline content*), wenn sie sich auf Inhalte beziehen, die nach den Community-Richtlinien an der Schwelle des Unzulässigen stehen.

Einleitung

Seit Mitte der 2000er Jahre entwickeln große Tech-Unternehmen wie Facebook (Meta), Google oder Netflix Ranking-Algorithmen, die darüber entscheiden, welche Inhalte den Nutzer:innen auf ihren jeweiligen Plattformen empfohlen werden. Dabei ergab sich die Notwendigkeit für eine gewisse Form der Sortierung von Inhalten zunächst allein aus deren zunehmender Menge. Heute sind die meisten Ranking-Algorithmen so konzipiert, dass sie den Nutzer:innen Inhalte prsentieren, um bestimmte Unternehmensmetriken zu maximieren. Als solche zielen sie beispielsweise darauf ab, die Interaktionsbereitschaft sowie die Verweildauer der Nutzer:innen auf einer Plattform oder die Anzahl der Aufrufe eines Videos zu beeinflussen.

Inzwischen liegt eine wachsende Anzahl von – zumeist qualitativen – Studien vor, die auf die negativen Auswirkungen von Empfehlungsalgorithmen auf die Nutzer:innen und den gesellschaftlichen Diskurs hinweisen.[3] Um vor diesem Hintergrund die Sicherheit der Nutzer:innen zu gewährleisten und demokratischen Austausch zu schützen, erwägen viele Regierungen zunehmend, regulierend in die Ausgestaltung der Algorithmen einzugreifen.

Große Aufmerksamkeit erregten in diesem Zusammenhang die Anschuldigungen der Whistleblowerin Frances Haugen, die nach ihrem Ausscheiden als Produktmanagerin bei Facebook 2021 interne Dokumente ihres ehemaligen Arbeitgebers offengelegt hatte. Diese lieferten Anhaltspunkte dafür, wie gezielte unternehmerische Entscheidungen die Nutzerinteraktionen steigern sollten und dazu führten, dass Algorithmen negative, toxische oder hasserfüllte Inhalte priorisierten und verstärkten. Mitte 2022 geriet das Thema Empfehlungsalgorithmen erneut ins Schlaglicht der Öffentlichkeit, als sich Macro-Influencer:innen wie Kim Kardashian und Kylie Jenner einer Kampagne anschlossen, die sich gegen Änderungen der Instagram-Algorithmen richtete. Diese Änderungen führten vor allem dazu, dass die Feeds der Nutzer:innen mit Videobeiträgen von nicht-abonnierten Accounts überschwemmt wurden. Eine Petition mit dem Titel »Make Instagram Instagram again« auf der Website Change.org sammelte mehr als 300.000 Unterschriften, um gegen die »TikTokisierung« von Instagram zu protestieren. Die Petition fordert eine Rückkehr zu chronologischen Feeds, gleichzeitig einen Algorithmus, »der Fotos gegenüber Videos favorisiert«, sowie mehr Wertschätzung für Content Creator:innen.[4]

Wenngleich die erwähnten Forderungen nicht unbedingt auf eine Lösung der zugrundeliegenden Probleme abzielen, stehen hinter dem Aufschrei der Nutzer:innen und Creator:innen die gleichen Bedenken, die auch in der politischen und wissenschaftlichen Debatte geäußert werden: Macht-Ungleichgewichte in Bezug auf die Entscheidungs- und Gestaltungsprozesse des algorithmischen Rankings.

Im Sommer 2022 erklärte Mark Zuckerberg (CEO von Meta) in einer Telefonkonferenz, die aktuell »wichtigsten Veränderungen« des Geschäftsmodells des Unternehmens bestünden darin, »dass die sozialen Feeds nicht mehr in erster Linie von den Personen und Accounts bestimmt werden, denen man folgt, sondern zunehmend auch von der KI, die einem Inhalte empfiehlt, die auf Facebook oder Instagram interessant sein könnten, unabhängig davon, ob man den Urhebern der Inhalte folgt oder nicht«. David Wehner (CFO von Meta) erklärte überdies, dass die »explizite Strategie des Unternehmens darin bestand, Reels [ein Kurzvideoformat] für eine größere Anzahl an Nutzer:innen sichtbar zu machen«.[5] Damit reagierte Meta auf Wettbewerber wie das chinesische Unternehmen TikTok und griff seinerseits den Trend zu kurzen Videos im Hochformat auf. Das heißt allerdings noch lange nicht, dass auch die Nutzer:innen Reels als Inhaltsformat und Plattformfunktionalität bewusst oder tatsächlich bevorzugen. Vielmehr sind es die Unternehmensmetriken, die den Content Creator:innen einen Anreiz bieten, bevorzugt Videos zu posten, um im Ranking-Feed empfohlen zu werden (wobei es durchaus möglich ist, dass einige Nutzer:innen Videos aus anderen Gründen gegenüber Fotos bevorzugen).

Mit der zunehmenden Bedeutung von Algorithmen lässt sich nunmehr feststellen, dass viele Content Creator:innen bei der Veröffentlichung von Videos »offen und aggressiv auf Viralität setzen«.[6] Eine ähnliche Dynamik entfaltet sich jedoch auch, wenn grenzwertige

oder schädliche Inhalte »viral gehen«, also eine sehr hohe Anzahl von Nutzer:innen erreichen. Diese Gefahr besteht insbesondere bei Kurzvideoformaten, deren »virale« Ausbreitung sich durch aktuell eingesetzte Prozesse zur Inhaltsmoderation kaum stoppen lässt, da letztere für andere Formate und Produkte konzipiert wurden.[7]

Auch wenn manche Creator:innen behaupten, die Algorithmen »überlisten« zu können, gibt es nach wie vor nur sehr wenig Transparenz und Überprüfbarkeit algorithmischer Ranking-Systeme. Eine fundierte Wissensbasis der (beabsichtigten wie unbeabsichtigten) Auswirkungen dieser Algorithmen nicht nur auf den Einzelnen, sondern auch auf eine breitere gesellschaftliche Wahrnehmung ist unverzichtbare Voraussetzung dafür, beim Entwickeln und Testen von Ranking-Systemen Verantwortung und Sorgfalt walten lassen zu können. Trotz (oder gerade wegen) der begrenzten Verfügbarkeit von Beweisen für die gesellschaftlichen Auswirkungen des algorithmischen Rankings wird in diesem Bericht versucht, die derzeitigen Ranking-Praktiken zu erläutern, Methoden zur Auditierung von Algorithmen zu untersuchen und über mögliche alternative Interventionen nachzudenken.

Abschnitt 1: Ranking-Algorithmen

1.1 Grundlagen algorithmischer Ranking-Systeme

Der von Facebook eingeführte algorithmisch generierte News Feed läutete eine allgemeine Abkehr von chronologischen Feeds ein, die in den Frühzeiten der sozialen Netzwerke üblich waren[8] – unterdessen gibt es Plattformen wie YouTube oder TikTok, für die chronologische Feeds kein offensichtliches »Standard«-Ranking darstellt. Vor diesem Hintergrund befasst sich dieser Abschnitt in erster Linie mit dem Beispiel des News Feeds und seinen zentralen Metriken. Die Betrachtungen sollen zum besseren Verständnis aufkommender algorithmischer Phänomene und deren potenziellen Einfluss auf den Informationskonsum und das Verhalten der Nutzer:innen beitragen.

2018 kündigte Mark Zuckerberg für den News Feed eine grundlegende Veränderung an. Mit den Änderungen sollten gemäß Wortlaut des Unternehmens »bedeutsame Interaktionen zwischen den Nutzer:innen (*Meaningful Social Interactions, MSI*)« gefördert werden.[9] Laut Meta sollten die Algorithmen des News Feeds Beiträge von Freunden und Familienmitgliedern, also der eigenen Community, gegenüber öffentlich zugänglichen Inhalten bevorzugen, und zwar insbesondere Beiträge, »die zu Diskussionen in den Kommentaren anregen« und »die dazu anregen, sie zu teilen und auf sie zu reagieren«.[10] Im vorliegenden Abschnitt werden zunächst die wichtigsten Komponenten des algorithmischen Rankings erläutert, damit im Anschluss die Auswirkungen des auf MSI basierenden Rankings beleuchtet werden können.

Algorithmische Ranking-Systeme der Plattformen bewerten Inhalte üblicherweise nach dem »zu erwartenden Engagement« und weisen ihnen auf dieser Grundlage ein entsprechendes Ranking zu. Inhalte jeder Art durchlaufen dabei einen Zyklus von der Inventarisierung und Charakterisierung über die Auswertung durch Modelle des Maschinellen Lernens (ML) bis hin zur Aufstellung von Prognosen und von Gesamtpunktzahlen. Alle genannten Prozesse werden

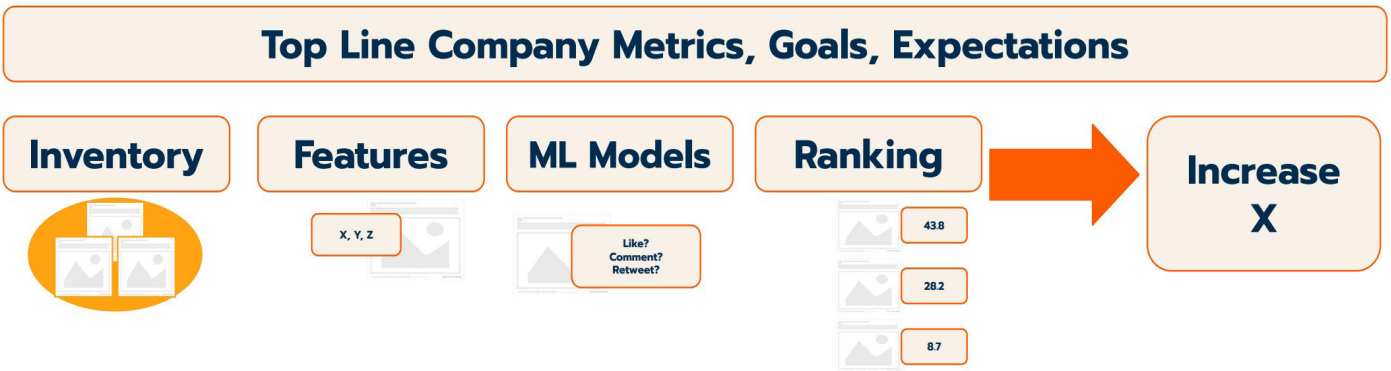

Abbildung 1: Übergeordnete Metriken, Ziele und Erwartungen von Unternehmen. Darstellung vom Integrity Institute.

von übergeordneten Unternehmensmetriken bestimmt. Ranking-Algorithmen inventarisieren zunächst alle verwertbaren Inhalte auf der Plattform, einschließlich öffentlicher Inhalte von nicht-abonnierten Accounts. **Anschließend greifen die Algorithmen auf sogenannte Ranking-Signale zurück, die Berechnungen von Millionen von Datenpunkten über den Verlauf, die Sichtbarkeit, die Merkmale und die Verteilung des Nutzer:innenverhaltens und der Inhalte umfassen.** Dabei werden beispielsweise folgende Faktoren berücksichtigt:

- Hat der betreffende Nutzer bereits ähnliche Inhalte »gelikt«, also mit »Gefällt mir« markiert?
- Welche Beziehung besteht zwischen Betrachter:in und Verfasser:in eines Beitrags (z. B.: Stammt der Inhalt aus dem Freundeskreis des Nutzers oder der Nutzerin oder ist es ein öffentlicher Beitrag)?
- Um welche Art von Beitrag handelt es sich (Video, Audio, Text, Bild oder eine Kombination)?

Umfangreiche statistische ML-Modelle verwenden die Signale zur Berechnung von Wahrscheinlichkeiten. Dabei beantworten sie Fragen wie z. B. »Wenn bestimmte Nutzer:innen diesen Inhalt schon einmal gesehen haben, wie hoch ist die Wahrscheinlichkeit, dass er ihnen gefällt, sie ihn kommentieren, teilen oder anschauen?« So wird Nutzer:innen, die zuvor bereits mit dem Verfasser eines Beitrags interagiert haben, beispielsweise eine höhere Wahrscheinlichkeit dafür zugewiesen, sich vermutlich erneut mit einem Beitrag desselben Verfassers zu beschäftigen. Falls eine bestimmte Nutzerin tendenziell ein höheres Engagement mit Videos eingeht, setzen die ML-Modelle eine höhere Wahrscheinlichkeit dafür an, dass sich diese Nutzerin mit Beiträgen in diesem Format beschäftigt, und diese beispielsweise mit »Gefällt mir« markiert oder kommentiert. Ranking-Algorithmen berechnen auch Wahrscheinlichkeiten für Fragen wie »Wird eine Nutzerin einer Seite folgen oder sich mit einem anderen Nutzer anfreunden, der mit dem jeweiligen Inhalt interagiert?« Bei einer Optimierung hin zu MSI werden Inhalte danach gewichtet, was die Plattform als »bedeutsame Konversationen« (*meaningful conversations*) einstuft. So verwenden die Algorithmen unterschiedliche Gewichtungen für verschiedene vorhergesagte Aktionen. Ein »Gefällt mir« wird so mit einem Punkt, eine »Reaktion oder ein Reshare ohne Text« mit fünf Punkten und »bedeutende Kommentare oder Reshares« mit 30 Punkten gewichtet.[11]

Abschließend berechnen ML-Modelle ein Ranking für alle Inhalte, die für einen bestimmten Nutzer verfügbar sind. Alle Punktzahlen werden kombiniert, um ein endgültiges Ranking zu ermitteln, mit dem nahezu in Echtzeit für alle Nutzer:innen ein Feed generiert und angezeigt werden kann. Beim endgültigen Ranking führt der News Feed einen »kontextbezogenen Durchlauf« durch, um eine Vielfalt von Formaten zu gewährleisten.[12] Während Ranking-Teams beschließen können, bestimmte Metriken zu maximieren, legt die Unternehmensführung die übergeordneten Metriken fest, um die Wirksamkeit des Ranking-Algorithmus zu bewerten – beispielsweise anhand der Anzahl der täglich aktiven Nutzer:innen, der Gesamtzahl der Beiträge oder der durchschnittlichen Nutzungsdauer auf der Plattform. Algorithmisches Ranking wird keinesfalls nur für Feeds verwendet, sondern auch für eine ganze Reihe weiterer Plattformfunktionen. Wichtige Metriken von YouTube sind zum Beispiel die Klicks (Anzahl der Aufrufe eines Videos), die Wiedergabezeit (Dauer in Minuten, die Nutzer:innen mit dem Betrachten des Videos verbringen) und das Engagement (Kommentare, Likes, Shares, usw.).[13] Auch wenn dieser Abschnitt den News Feed fokussiert, nutzen Plattformen algorithmische Ranking-Systeme für verschiedene Inhaltsformate und Funktionalitäten, beispielsweise vorgeschlagene Accounts, Push-Mitteilungen oder die automatische Vervollständigung von Suchfunktionen.

1.2 Engagement-Problem, Superuser und andere algorithmische Phänomene

Ein zentrales Problem bei Ranking-Systemen, die auf eine Steigerung des Engagements abzielen, ist das »natürliche Interaktionsmuster« (*natural engagement pattern*). So steigt die Bereitschaft zum Engagement (also zu Interaktionen wie »liken«, teilen, kommentieren), wenn Inhalte sich einem Grenzwert dessen nähern, was auf einer Plattform erlaubt ist. Meta definiert diese Inhalte als »grenzwertig, bezogen auf die Facebook-Gemeinschaftsstandards«, wenn sie nach den Gemeinschaftsstandards nicht verboten sind, aber nahe an die Grenzen dieser Richtlinien herankommen. Dazu zählen Verhaltensweisen, die an Mobbing und Belästigung, Hassrede, Gewalt und Aufwiegelung grenzen, oder Inhalte, die irreführend oder reißerisch sind.[14]

Daraus ergibt sich das in der Abbildung 2 veranschaulichte »Engagement-Problem« (x-Achse: Bereiche mit zulässigen und verbotenen Inhalten; y-Achse: Maß für das Engagement). 2018 erklärte Mark Zuckerberg: »Unsere Forschung zeigt auf, dass die Menschen sich im Durchschnitt mehr mit einem Inhalt beschäftigen, wenn er sich der Grenze des Erlaubten nähert – unabhängig davon, wo wir diese Grenzen ziehen und selbst wenn sie uns im Nachhinein sagen, dass sie den Inhalt nicht mögen«. Ein ähnliches Eingeständnis ist von dem ehemaligen YouTube-Softwareentwickler Guillaume Chaslot bekannt. Er machte seine Kolleg:innen darauf aufmerksam, dass auf der Videoplattform verschwörungsideologische Inhalte des von Alex Jones betriebenen Online-Portals InfoWars empfohlen wurden, worauf jemand antwortete: »Die Leute klicken halt drauf. Was sollen wir da tun?«[15]

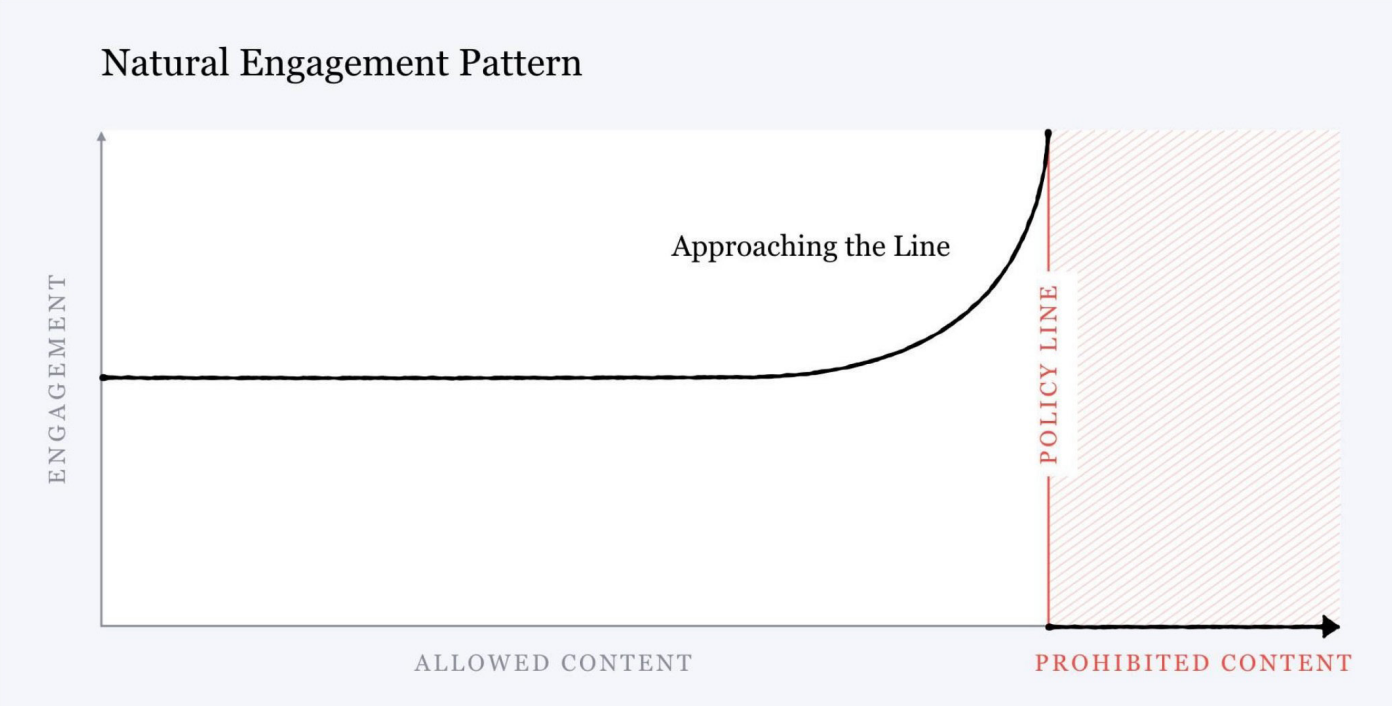

Abbildung 2: Natürliches Schema, nach dem grenzwertige Inhalte die meisten Interaktionen (Engagement) erzielen. Darstellung vom Integrity Institute.

Ein auf Engagement basierendes Ranking bevorzugt Inhalte, die sich nahe an dieser Grenzlinie befinden, da Inhalte, von denen man sich mehr Engagement erwartet, mit größerer Wahrscheinlichkeit schädlich sind. Die zunehmende Tendenz zu grenzwertigen und schädlichen Inhalten wird zum Problem. Einerseits stufen ML-Modelle ein Nutzer:innenverhalten, welches das Engagement maximiert, hoch ein. Ein solches Verhalten verstößt aber in manchen Fällen gegen die Community-Richtlinien. Die Ausrichtung auf das Erreichen von Metriken kann daher schnell dazu führen, dass das algorithmische Ranking auf Engagement und damit auf bösartige, sensationalistische und grenzwertige Inhalte optimiert wird.

Das hat Konsequenzen für die reale Welt. Aus einer internen Aktennotiz vom April 2019 geht beispielsweise hervor, dass sich politische Parteien in Europa aufgrund der Ranking-Algorithmen von Facebook gezwungen sahen, »erheblich mehr negative Inhalte als zuvor« zu veröffentlichen, da das Engagement bei positiven Beiträgen bzw. Beiträgen zu politischen Maßnahmen dramatisch zurückgegangen sei. Dies liege daran, dass der MSI-Mechanismus »systematisch provokative, minderwertige Inhalte« belohne.[16] Die Unternehmensmetriken belohnten Nutzer:innen oder Gruppen, die spalterische, schockierende oder irreführende Inhalte veröffentlichten. Damit verbunden ist das Risiko, dass für Inhaltserstellende wie politische Parteien oder Politiker:innen wenig Anreiz besteht, differenziertere und faktenbasierte Informationen zu veröffentlichen.

Ein auf Engagement ausgerichtetes Ranking birgt darüber hinaus die Gefahr eines ungewollten Macht-Ungleichgewichts zwischen den Nutzer:innen selbst. Im Sommer 2020 untersuchten Forscher das Phänomen einer Vormachtstellung sogenannter Superuser auf Facebook. Gemeint ist eine Gruppe von Nutzer:innen, die mehr Likes, Shares, Reaktionen, Kommentare und Posts produzieren als 99 Prozent aller Nutzer:innen in den USA. Die Forscher analysierten 52 Millionen Nutzer:innen und 500 in den USA betriebene Seiten mit dem höchsten durchschnittlichen Engagement sowie die Beiträge mit der höchsten Interaktion aus mehr als 41.000 der öffentlichen Gruppen mit der höchsten Mitgliederzahl in den USA. Dabei stellten sie fest, dass ein Prozent der Accounts für 35 Prozent aller registrierten Interaktionen verantwortlich war; die aktivsten drei Prozent waren für 52 Prozent verantwortlich. Viele Nutzer:innen interagierten selten oder gar nicht mit öffentlichen Gruppen oder Seiten. Die Untersuchungen ergaben, dass diese einflussreichen Nutzer:innen auch am häufigsten missbräuchlich handelten und das öffentlich zugängliche Inventar so in Richtung grenzwertiger Inhalte verschieben. Bei einer zufällig ausgewählten Stichprobe von 30.000 Nutzer:innen, die sich auf die 219 Accounts mit mindestens 25 öffentlichen Kommentaren bezog, fiel auf, dass 68 Prozent Fehlinformationen verbreiteten, spambehaftete Inhalte sowie rassistische, sexistische, antisemitische oder schwulenfeindliche Kommentare veröffentlichten oder zu Gewalt aufhetzten.[17] Weil Ranking-Algorithmen »sinnvolle Interaktionen« (*meaningful interactions*) wie Kommentare besonders belohnen, hätten Superuser einen unverhältnismäßig großen Einfluss darauf, wie die Algorithmen ermitteln, was für andere Nutzer:innen interessant sein könnte. Den Sackgasseneffekt, der dabei entsteht, nennt Sahar Massachi vom Integrity Institute eine »stakeholder trap«: Wenn eine Plattform die »falschen« Nutzer:innen lange genug belohnt, lässt sie zu, dass diese Nutzer:innen sehr mächtig werden und die Plattformen und ihre Algorithmen in eine Sackgasse führen.[18]

Das Problem der Machtstellung von Superusern weist gleichzeitig auf eine weitere Schwäche von algorithmischen Ranking-Praktiken hin, nämlich die mangelnde Gleichberechtigung aller Nutzer:innen. Einschlägige Untersuchungen zeigen, dass eingesetzte ML-Algorithmen häufig gewisse Vorurteile fördern, von denen bereits marginalisierte Gruppen potenziell überproportional betroffen sind. So untersuchten Wissenschaftler:innen beispielsweise rassistische Tendenzen in Datensätzen von Twitter zur Erkennung von Hassrede und beleidigender Sprache. Die Untersuchungen ergaben Hinweise auf eine systematische rassistische Benachteiligung, da die Klassifizierungssysteme dazu neigen würden, für in afroamerikanischem Englisch verfasste Tweets eine wesentlich höhere Wahrscheinlichkeit für Missbrauch vorherzusagen.[19] Eine weitere vielbeachtete Arbeit von Dr. Safiya Umoja Noble aus dem Jahr 2018 wies zudem nach, dass Algorithmen von Online-Suchmaschinen gesellschaftliche Vorurteile festigen, insbesondere die Diskriminierung von Schwarzen Menschen und allen voran Schwarzen Frauen.[20]

Ranking-Algorithmen greifen heute auf ein umfangreiches Inventar an Inhalten zurück, die zu einem gewissen Teil rassistisch, sexistisch oder auf sonstige Weise diskriminierend sind. Die Möglichkeiten für menschliche Aufsicht und Kontrolle sind dagegen unzureichend. Die an den Rankings arbeitenden Teams der (zumeist) in den USA ansässigen Plattformen sind oft nicht in der Lage, die Auswirkungen auf die Inhalte und das Verhalten der Nutzer:innen unter sehr unterschiedlichen geografischen, sprachlichen oder rechtlichen Rahmenbedingungen abzuschätzen[21], was die Ungleichbehandlung und Diskriminierung der Nutzer:innen zusätzlich verstärkt. **Zwar sind systematische Verzerrungen in Verbindung mit dem Problem eines auf Engagement zielenden Rankings kein neues Phänomen. Allerdings können sie diskriminierendes Verhalten und menschliche kognitive Verzerrungen verstärken.** Auf ähnliche Weise haben geschäftliche Interessen auch in der Vergangenheit kognitionspsychologische Phänomene verschärft – so bei traditionellen Massenmedien, insbesondere in Boulevardzeitungen durch sensationalistische Schlagzeilen oder Framings. **Aviv Ovadya vom Belfer Center an der Harvard Kennedy School weist zudem darauf hin, dass in diesem Zusammenhang auch evolutionäre Gründe eine Rolle spielen könnten und schlussfolgert, dass »Menschen nicht immer faktenbasiert denken und unabhängig vom zugrundeliegenden Wahrheitsgehalt sehr auf Sensationalismus und Spalterei anspringen«.[22]**

Algorithmische Ranking-Praktiken befassen sich im Kern mit der Frage »wer kontrolliert, was wir sehen und wie wir es sehen«[23] und stellen ungleiche Machtverhältnisse zwischen Plattformen und Nutzer:innen dar. Dabei zielen die Unternehmensmetriken, die den »Erfolg« der Ranking-Systeme messen, vor allem auf erhöhtes Engagement seitens der Nutzer:innen ab, was wiederum grenzwertigen Inhalten besonders viel Raum bietet. Einige der damit verbundenen gesellschaftlichen Risiken wurden in diesem Abschnitt dargestellt.

Abschnitt 2: Auditierung algorithmischer Ranking-Systeme

Anders als bei externen Forscher:innen besteht für die Mitarbeiter:innen von Tech-Unternehmen, die für die Entwicklung und das Testen von Ranking-Algorithmen verantwortlich sind, nicht notwendigerweise ein Anreiz, Risiken gesellschaftlicher Spaltungen und Verzerrungen aufzudecken. Mangels uneingeschränkten Zugriffs auf die Daten sind Forscher:innen wiederum auf die Unterstützung interner Mitarbeiter:innen angewiesen, um die Algorithmen zu untersuchen. Infolgedessen werden Ranking-Algorithmen und ihre tatsächlichen gesellschaftlichen Auswirkungen von unabhängigen Forscher:innen und Organisationen – und damit auch von politischen Entscheidungsträger:innen und der Öffentlichkeit – unzureichend verstanden.

Unabhängige Auditor:innen, die keiner vertraglichen Bindung mit den zu auditierenden Unternehmen unterliegen, könnten offen und unbefangen beurteilen, wie ein System funktioniert und wie etwaige Verzerrungen oder Risiken abgemildert werden können. Aufgrund fehlender klar geregelter Verfahren und Normen für die Audits gestalten sich die Durchführung fundierter Untersuchungen und die Überprüfung der Ergebnisse jedoch weiterhin schwierig. Dieser Abschnitt befasst sich mit den Möglichkeiten zur Durchführung von Audits der Ranking-Algorithmen durch Dritte und zeigt die Grenzen der bisherigen Methoden auf. Darüber hinaus wird in diesem Abschnitt auf die Notwendigkeit einer transparenten, klaren und standardisierten Auditpraxis zur Gewährleistung faktenbasierter politischer Maßnahmen und der Haftbarkeit der Plattformbetreiber eingegangen.

2.1 Audit-Methoden für Ranking-Algorithmen

Audits sind ein zentrales Instrument für politische Interventionen, die auf die Schaffung einer evidenzbasierten Grundlage und die Einhaltung von Sorgfaltspflichten abzielen. KI-gestützte Technologien durchdringen immer mehr Aspekte der Gesellschaft. Sie werden beispielsweise unterstützend eingesetzt, um in großen Unternehmen über Einstellungen und Entlassungen oder die Gewährung von Darlehen zu entscheiden. Dementsprechend werden die ethischen Auswirkungen dieser Systeme in verschiedenen Politikfeldern zunehmend diskutiert.[24] Die Auditierung solcher Algorithmen kann der allgemeinen Überwachung, der Untersuchung eines bestimmten Schadens oder Risikos oder der Durchsetzung gesetzlicher Pflichten dienen. So konzentrieren sich Audits bei Algorithmen, die Bewertungen von Menschen vornehmen, auf Hinweise für eine Diskriminierung bestimmter Gruppen durch potenzielle Verzerrungen, deren Ursache im ML-Modell selbst oder in gelabelten Daten liegen, die für das Training der Modelle eingesetzt wurden. Eine Überprüfung kommerzieller automatischer Algorithmen und Datensätze zur Gesichtserkennung durch die Wissenschaftlerinnen Joy Buolamwini und Timnit Gebru ergab, dass das System Menschen mit dunkler Hautfarbe weniger gut erkennt als Menschen mit heller Hautfarbe.[25]

Im Zusammenhang mit dem algorithmischen Ranking der Social-Media-Plattformen sind aus der Forschung unterschiedliche Vorschläge für Methoden bekannt, mit denen bewertet werden soll, wie Systeme den Online-Diskurs und das Verhalten der Nutzer:innen beeinflussen. Jede dieser Methoden hat eigene Vor- und Nachteile. Dabei ist die Vielfalt der methodischen Ansätze vor allem der Tatsache geschuldet, dass ein umfassender und einheitlicher Zugriff auf interne Daten und Forschungsexperimente der Plattformen fehlt. Die nachstehend dargestellte Übersicht wurde vom Ada Lovelace Institute erarbeitet. Darin werden der Zweck und die Nachteile der einzelnen Methoden zur regulatorischen Inspektion von algorithmischen Ranking-Systemen vergleichend gegenübergestellt.[26]

Auditmethode	Beschreibung	Zweck	Nachteile
Code-Audit	Auditor:innen haben direkten Zugriff auf die Codebasis des betreffenden Systems oder auf den »Pseudocode«, d. h. auf eine verständliche Veranschaulichung der Funktionsweise des Codes für den Algorithmus.	Verständnis der mit den Algorithmen verfolgten Absichten; im Falle von Maschinellem Lernen hilfreich, um zu verstehen, welche Ziele die Optimierung bestimmen.	Die Codebasis kann sehr groß sein – einzelne Softwareentwickler:innen in großen Unternehmen wissen selten, wie alle Teile der Plattform funktionieren. Die Auswirkungen lassen sich anhand des Codes nur schwer erkennen. Bedenken hinsichtlich Verletzungen der Rechte und des Schutzes des
Befragung der Nutzer:innen	Die Auditor:innen führen Umfragen und/oder Einzelinterviews im Kreis der Nutzer:innen durch, um deskriptive Daten über deren Erfahrungen auf einer bestimmten Plattform zu erfassen.	Zusammentragen von Informationen über die Nutzererfahrungen auf einer Plattform zur groben Identifizierung problematischer Verhaltensmuster, die anschließend einer genaueren Untersuchung zugeführt werden.	Anfällig für die typischen sozialwissenschaftlichen Bedenken bei Umfragen – Druck, auf eine bestimmte Art und Weise antworten zu müssen, Unzuverlässigkeit des menschlichen Gedächtnisses und die Schwierigkeit, den Ergebnissen einen kausalen
Scraping-Audit (automatisiertes Auslesen)	Auditor:innen sammeln Daten direkt von einer Plattform, in der Regel, indem sie einen Code schreiben, der automatisch auf eine Website klickt oder diese durchsucht, um relevante Daten zu sammeln (z. B. von den Nutzer:innen veröffentlichter Text).	Verständnis des Inhalts, der auf einer Plattform dargestellt wird, insbesondere durch beschreibende Aussagen (z. B. »Dieser Anteil der Suchergebnisse enthielt diesen Begriff«) oder durch Vergleich von Ergebnissen für unterschiedliche Gruppen oder Begriffe.	Erfordert die Entwicklung eines spezifischen Tools für jede Social-Media-Plattform; dies kann problematisch sein, da bereits kleine (legitime) Änderungen am Layout einer Website das Programm unbrauchbar machen können.
API-Audit	Die Auditor:innen greifen über eine plattformseitig bereitgestellte Programmierschnittstelle auf die Daten zu, die es ihnen ermöglicht, über selbst geschriebene Computerprogramme Informationen an die Plattform zu senden und von der Plattform zu empfangen; Beispiel: eine API kann es einem Nutzer ermöglichen, ein Schlüsselwort zu senden und eine Anzahl an Treffern zurück zu erhalten.	Im Vergleich zum Scraping-Audit ist der programmiertechnische Zugang zu den Daten vereinfacht; dies ermöglicht eine unkompliziertere Automatisierung der Datenerfassung für beschreibende Aussagen oder vergleichende Zwecke.	Öffentlich zugängliche APIs liefern Aufsichtsbehörden möglicherweise nicht die tatsächlich benötigten Daten. Mit entsprechenden Befugnissen zur Erfassung von Informationen könnten sie Plattformbetreiber stattdessen dazu verpflichten, Zugang zu weiteren APIs oder sogar zu einer spezifischen (eigens entwickelten) API zu gewähren; dies würde jedoch unter Umständen zusätzlichen Entwicklungsaufwand auf Seiten der Plattformbetreiber erfordern.

Auditmethode	Beschreibung	Zweck	Nachteile
Sock puppet-Audit	Die Auditor:innen geben sich durch Nutzung entsprechender Computerprogramme als Nutzer:innen der Plattform aus. Die von der Plattform als Reaktion auf die programmierten Accounts erzeugten Daten werden aufgezeichnet und ausgewertet.	Einblick in mögliche Erfahrungen bestimmter Accounts oder Nutzer:innengruppen auf der Plattform.	Sock puppet Accounts täuschen Nutzer:innen nur vor – sie sind keine echten Nutzer:innen und insofern bestenfalls Stellvertreter:innen für individuelle Aktivitäten und Erfahrungen echter Nutzer:innen. Gleichwohl haben sich Sock puppet-Audits für Forschungszwecke als sinnvoll erwiesen, um ein grundlegendes Verständnis über das
Crowdsourcing	Bei einem Audit durch Crowdsourcing sammeln echte Nutzer:innen im Zuge ihrer Plattformnutzung Informationen. Dies kann durch manuell erstellte Erfahrungsberichte oder anhand von automatisierten Mechanismen wie Browsererweiterungen erfolgen.	Beobachtungen in Bezug auf den Inhalt, der den Nutzer:innen auf den Plattformen ausgespielt wird, einschließlich unter der Fragestellung, ob unterschiedlichen Accounts unterschiedliche Inhalte ausgespielt werden.	Erfordert einen individuellen, häufig auf Web-Scraping-Techniken beruhenden Ansatz zur Datenerhebung für jede zu auditierende Social-Media-Plattform; da die Methode bisher nur für Desktop- und nicht für Mobilgeräte angewendet wurde, könnten für Mobilgeräte spezifische Ergebnisse abweichen oder Erfahrungen unberücksichtigt bleiben.

Tabelle übernommen vom Ada Lovelace Institute.

Policy Initiative: Digital Services Act (DSA) der EU

Mit dem Gesetz über digitale Dienste (Digital Services Act, DSA), das am 27. Oktober 2022 unterzeichnet und im Amtsblatt der Europäischen Union veröffentlicht wurde[27], werden Anbieter von sehr großen Online-Plattformen (VLOPs) bzw. sehr großen Online-Suchmaschinen (VLOSEs)[I] zur Einhaltung einer Reihe einschlägiger Maßnahmen in Bezug auf Transparenz und Sorgfaltspflicht verpflichtet. Durch die Verordnung werden mehrere Prüfebenen eingeführt: verbindliche interne Risikobewertungen (Eigenaudits), verbindliche externe Prüfungen durch unabhängige Sachverständige, die die Einhaltung der Verpflichtungen (einschließlich in Bezug auf die Risikobewertungen) überprüfen, sowie unabhängige Prüfungen durch zugelassene Forscher:innen im Rahmen der Bestimmungen für den Datenzugang.

Im Rahmen der verbindlichen Risikobewertungen (Artikel 34) sind Plattformen dazu verpflichtet, alle systemischen Risiken zu ermitteln, zu analysieren und zu bewerten, die sich aus der Ausgestaltung oder dem Betrieb ihrer Dienste und der damit verbundenen Systeme, einschließlich algorithmischer Systeme, ergeben. Dies umfasst die systemischen Risiken in Verbindung mit der Verbreitung rechtswidriger Inhalte, tatsächliche oder absehbare negative Auswirkungen auf die Ausübung der Grundrechte, auf die gesellschaftliche Debatte, auf Wahlverfahren und auf die öffentliche Sicherheit oder in Bezug auf geschlechtsspezifische Gewalt, den Schutz der öffentlichen Gesundheit und von Minderjährigen sowie schwerwiegende negative Folgen für das körperliche und geistige Wohlbefinden von Personen. Besonders hervorzuheben ist, dass die Plattformen dabei die Ausgestaltung ihrer algorithmischen Ranking-Systeme berücksichtigen müssen. Wenn die algorithmische Verstärkung von Informationen zu systemischen Risiken beiträgt, müssen die Anbieter dies in ihre Risikobewertungen und damit auch in ihre Maßnahmen zur Risikominderung einbeziehen (Artikel 35).

Sowohl die Risikobewertungen als auch die Risikominderungsmaßnahmen sind jährlichen unabhängigen Prüfungen zu unterziehen. Die Europäische Kommission kann darüber hinaus delegierte Rechtsakte für spezifischere Vorschriften erlassen, zum Beispiel in Bezug auf die Methoden für die Prüfung (Auditierung) und Vorlagen für die Berichterstattung (Artikel 37). In diesem Zusammenhang wird das European Centre for Algorithmic Transparency (ECAT), das in enger Zusammenarbeit mit dem Directorate-General for Communications Networks, Content and Technology (DG CONNECT) und dem Joint Research Centre betrieben wird, »wissenschaftliches und technisches Fachwissen« zur ausschließlichen Aufsichts- und Durchsetzungsfunktion der Kommission beitragen und technische Unterstützung wie »technische Tests algorithmischer Systeme« oder wissenschaftliche Forschung wie »praktische Methoden für transparente und überprüfbare algorithmische Ansätze« bereitstellen.[28]

Zum Zwecke der Überwachung der Konformität mit den sich aus der Verordnung ergebenden Pflichten können die nationalen Aufsichtsbehörden der EU-Mitgliedstaaten oder die Europäische Kommission den Zugang zu oder die Weitergabe von bestimmten Daten verlangen, darunter auch Informationen über die »Gestaltung, die Logik, die Funktionsweise und die Tests ihrer algorithmischen Systeme«. Auf Anfrage der Aufsichtsbehörden müssen die Plattformen auch zugelassenen Forscher:innen den Zugang zu Daten gewähren, die zur »Aufspürung, zur Ermittlung und zum Verständnis systemischer Risiken beitragen, auch in Bezug auf die Bewertung der Angemessenheit, der Wirksamkeit und der Auswirkungen der Risikominderungsmaßnahmen« (Artikel 40). Diesbezüglich weist Mathias Vermeulen darauf hin, dass zugelassene Forscher:innen (*vetted researchers*) damit im Prinzip als Auditor:innen betrachtet werden könnten: Sie können aufkommende Risiken bewerten, die möglicherweise nicht durch den internen Risikobewertungsbericht einer Plattform abgedeckt wurden, und beurteilen, ob die Maßnahmen zur Risikominderung in der Praxis wirksam sind.[29]

[I] Zu den VLOPs und VLOSEs werden Online-Plattform- und Suchmaschinenanbieter gezählt, deren Zahl aktiver Nutzer:innen die operative Schwelle von 45 Millionen – 10 % der Bevölkerung in der Union – überschreitet (im Durchschnitt über einen Zeitraum von sechs Monaten).

Darüber hinaus sollen freiwillige Verhaltenskodizes zur ordnungsgemäßen Anwendung des DSA beitragen, unter anderem durch Festlegung von Verpflichtungen zur Ergreifung spezifischer Risikominderungsmaßnahmen sowie eines Rahmens für die regelmäßige Berichterstattung über alle ergriffenen Maßnahmen und deren Ergebnisse (Artikel 45). Der im Juni 2022 verschärfte Verhaltenskodex zur Bekämpfung von Desinformation (*Code of Practice on Disinformation CoPD*)[30] verpflichtet die Unterzeichner, zu denen Google (YouTube), Meta (Facebook, Instagram), Microsoft (LinkedIn), TikTok und Twitter gehören, zur Minimierung der Risiken einer »viralen Verbreitung von Desinformation« durch die Anwendung sicherer Praktiken zur Ausgestaltung ihrer jeweiligen Dienste. Die Unternehmen verpflichten sich auch zu Investitionen in die Forschung über die Verbreitung von schädlichen Desinformationen im Internet und die damit verbundenen sicheren Praktiken zur Ausgestaltung der Dienste. Vermeulen merkt an, dass die Kommission zwar die Erfüllung dieser Verpflichtungen bei der Bewertung der Konformität mit dem DSA berücksichtigen werde. Die Datenzugangsregelung nach dem Verhaltenskodex diene jedoch einem anderen Zweck, da der Zugang zu jedem beliebigen Forschungszweck über »Desinformation«[31] gewährt werden könne und daher nicht zwangsläufig mit der Bewertung systemischer Risiken und Risikominderungsmaßnahmen in Verbindung stehen muss.[32]

2.2 Methodische und epistemische Limitationen

Vor dem Hintergrund variierender Datenzugangsvoraussetzungen entwickeln und verwenden unabhängige Forscher:innen unterschiedliche methodische Ansätze zur Erfassung und Auswertung von Plattformdaten. Trotz erster Forschungsergebnisse steht das Forschungsfeld weiterhin vor großen Hürden, wenn es darum geht, systematische, langfristige und weitreichende Datenanalysen durchzuführen, die ein umfassenderes Verständnis der Auswirkungen des algorithmischen Rankings auf den Online-Diskurs ermöglichen. Einerseits können Plattformbetreiber den Zugang zu Daten gezielt einschränken, wenn etwa Daten- oder Innovationsschutzgründe oder der Schutz von Geschäftsgeheimnissen dazu Anlass geben. Des Weiteren können auch durch technologische Merkmale unbeabsichtigte Barrieren entstehen (z. B. können Video- und Audioinhalte bislang nicht so einfach durchsucht oder ausgewertet werden wie Textinhalte). Der Zugang zu bestimmten Online-Bereichen, wie privaten und/oder verschlüsselten Nachrichten oder nichtöffentlichen Kanälen und Gruppen, kann und sollte auch aufgrund von praktischen und/oder ethischen Erwägungen eingeschränkt werden. In anderen Fällen können theoretisch zugängliche öffentliche Inhalte aufgrund fehlender Offenlegung oder fehlender Infrastrukturen für den Datenzugang wie APIs nicht ausreichend analysiert werden.[33]

Insgesamt besteht in Anbetracht des Spektrums und der Komplexität von ML-Algorithmen, die für ihre Berechnungen auf Millionen von Signalen zugreifen, eine Wissenslücke. Tatsächlich sind Ranking-Algorithmen schwer zu entschlüsseln und für externe Forscher:innen ohne Zugang zu den Experimenten, die für deren Entwicklung von den unternehmensinternen Teams durchgeführt werden, kaum zu beurteilen.

Die Abhängigkeit von Transparenz und Datenzugangsmöglichkeiten verstärkt die epistemischen Einschränkungen bei der Erforschung der Auswirkungen sozialer Medien auf gesellschaftliche Debatten. In Bezug auf das Problem der Datenverfügbarkeit bleibt den Forscher:innen bisher, sich auf bestimmte Plattformen zu konzentrieren, die einen vergleichsweise umfassenden Datenzugriff über API

bieten, was insbesondere auf Twitter zutrifft. Ein zu enger Betrachtungsrahmen führt jedoch dazu, dass bestimmte demografische oder geografische Gruppen über- oder unterrepräsentiert sind. So wird die Plattform TikTok[34] zum Beispiel überwiegend von einer jungen Zielgruppe genutzt, wobei die Funktionen der API nach wie vor sehr begrenzt sind und die Plattform nur sehr wenige sachdienliche Informationen über ihre Algorithmen preisgibt.[35] Andere Zielgruppen beziehen den Großteil ihrer Informationen über traditionelle Medien wie das Fernsehen, während wieder andere zur Befriedigung ihres Informationsbedarfs zunehmend Messaging-Anwendungen wie Telegram oder WhatsApp nutzen.[36]

Darüber hinaus fehlen klare Begriffsbestimmungen für einzelne Risiken und Maßnahmen zur Risikominderung, die bei den Audits analysiert und bewertet werden sollen. Ein Beispiel: Im Zusammenhang mit der Bundestagswahl 2021 beabsichtigte das Sustainable Computing Lab, eine externe Risikobewertung der systemischen Risiken für das Recht auf freie und faire Wahlen durchzuführen, und konzentrierte sich dabei auf Twitter und Facebook. Dabei mussten die Forscher:innen feststellen, dass sich Kategorien zur Identifizierung von Wahlrechtsverletzungen und Desinformation nur unter Schwierigkeiten gegeneinander abgrenzen lassen. Die Forscher:innen mahnten an, dass Kategorien wie »systemische Risiken« bei wahlbezogener Desinformation klarer abgegrenzt werden müssten, damit sie in zukünftigen Forschungsprojekten leichter reproduzierbar und vergleichbar sind. Noch wichtiger ist, dass politische Entscheidungsträger:innen und Forscher:innen häufig Begriffe wie »algorithmische Verstärkung« (*amplification*) oder »Herabstufung« (*demotion*) verwenden, ohne ein klares und nachvollziehbares Begriffsverständnis zu haben, wodurch eine Gefahr besteht, dass falsche Erwartungen in Bezug auf die Prüfungs- und Abhilfemaßnahmen entstehen.

Zusammenfassend lässt sich sagen, dass soziale Medien ein schnell wachsendes Forschungsfeld sind, das jedoch aufgrund der Einschränkungen methodischer Ansätze nach wie vor die »unknown unknowns« des Themenfeldes vor sich herschiebt. Dies führt nicht nur zu unterschiedlichen Standards hinsichtlich der Forschungskonzepte, sondern auch zu schlecht vergleichbaren Ergebnissen, deren Anwendungsbereich und analytischer Wert mitunter begrenzt sind. **Da die von den Gesetzgebern vorgeschlagenen verbindlichen Regeln vorsehen, dass Plattformbetreiber, Aufsichtsbehörden und externe Forscher:innen Prüfungen (Audits) durchführen, sind die Entwicklung eines transparenten Forschungskonzepts, die Verständigung auf klare Begriffsbestimmungen und eine robuste Methodik von entscheidender Bedeutung, um konsistente und wirksame Audits durch unabhängige Dritte zu ermöglichen.**

Qualitätsstandards und Transparenz algorithmischer Audits

Zivilgesellschaft, Wissenschaft, politische Entscheidungsträger:innen, Aufsichtsbehörden und die Plattformen selbst sollten offene und inklusive Konsultationen organisieren, um gemeinsame Standards und praxistaugliche Regeln für die Optimierung der Qualität und Transparenz bei der Prüfung von Algorithmen zu entwickeln. Die Prüfung von Algorithmen setzt die Entwicklung eines gemeinsamen epistemischen Verständnisses in Bezug auf Forschungsmethoden und Terminologie voraus. Um ein ausreichendes Maß an Legitimität zu gewährleisten, sollte dabei interdisziplinäres Fachwissen sowie demografische, geografische und linguistische Vielfalt eingebunden werden. Insbesondere erfordert eine wirksame Auditierung die Klärung der relevanten gesellschaftlichen Risiken und Maßnahmen zur Risikominderung, einschließlich in Bezug auf Änderungen der Metriken. Es wird darauf ankommen, praxistaugliche Definitionen für Konzepte wie »algorithmische Verstärkung« in überprüfbare Hypothesen zu überführen.[37] **Bei der Entwicklung von Methoden ist es wichtig zu betonen, dass es je nach Funktionalität einer Plattform sehr unterschiedliche Ausgangssituationen (*baselines*) gibt bzw. eine Ausgangssituation schwer zu definieren ist, wenn es um »algorithmische Verstärkung« geht (In anderen Worten: Im Vergleich zu welcher Ausgangssituation werden die Inhalte »verstärkt«?).**

Aus Sicht der Plattformen wird die Pflicht zur Durchführung interner Risikobewertungen die Zusammenarbeit mit externen Forscher:innen notwendig machen. Diese Zusammenarbeit bietet ihnen wiederum den Vorteil, dass die Teams bei der Ausgestaltung der Ranking-Mechanismen frühzeitig erkennen können, worauf sie achten sollten, damit die ML-Modellierung nicht zu einer Verschiebung in Richtung schädlicher Inhalte führt. Auf der anderen Seite werden externe Audits darauf angewiesen sein, dass die Plattformen ihre internen Entscheidungsprozesse in Bezug auf die Methodik und die Durchführung von Experimenten offenlegen. So könnten Plattformen beispielsweise zusammengefasste Statistiken offenlegen, die in den Entscheidungsprozessen des Unternehmens genutzt werden. Transparenz über interne Experimente - beispielsweise randomisierte Studien, die Nutzer:innen unterschiedliche algorithmische Ranking-Systeme zuweisen – könnte der Auditierung und Forschung wichtige Hinweise über die (unbeabsichtigten) Folgen der Ranking-Mechanismen geben. So könnten Plattformen beispielsweise Aufstellungen über Experimente mit entsprechenden Hypothesen, Daten und daraus abgeleiteten Entscheidungen offenlegen.[38] Ein solches Maß an aussagekräftiger Transparenz könnte externen Forscher:innen helfen, die Auswirkungen von Änderungen der Metriken sowie Ursache-Wirkung-Beziehungen zu verstehen.

Eine unabhängige Stelle, die als Vermittler zwischen Aufsichtsbehörden, Forschenden und Plattformen fungiert, wie im sich entwickelnden EU-Rechtsrahmen vorgesehen, könnte – nicht zuletzt auch in Anbetracht der Sensibilität der Daten von Algorithmen – bei der Entwicklung und Überwachung von Leitlinien für die Zulassungs- und Prüfverfahren, Datenschutz und Transparenz der Forschungsdokumentation unterstützen.

Die Möglichkeiten für die unabhängige Auditierung von Algorithmen könnten in transatlantischen und transpazifischen politischen Kreisen vorrangig behandelt werden, beispielsweise im Rahmen der Sitzungen des EU-US-Handels- und Technologierates (TTC) oder der Christchurch Call Community (wie z. B. durch die kürzlich gestartete *Initiative on Algorithmic Outcomes*[39]). Insbesondere könnte der EU-US-TTC die Koordinierung der verschiedenen Ansätze der geplanten Regulierungsmechanismen und Rahmenwerke für die KI (u. a. den Entwurf für die geplante KI-Grundrechtecharta des Weißen Hauses, die »AI Bill of Rights«[40]) voranbringen, und dabei auch koordinieren, welche Ansätze diese für die von Social-Media-Plattformen verwendeten algorithmischen Ranking-Systeme wählen.

Jegliche zwischenstaatliche Konsultation über algorithmische Ranking-Systeme sollte von verschiedenen Fachleuten und Perspektiven (einschließlich unter dem Aspekt der geografischen, geschlechtsspezifischen und sprachlichen Vielfalt), Vertreter:innen aus Nichtregierungsorganisationen und der Wissenschaft sowie den in den Unternehmen für die Ranking-Systeme zuständigen Teams begleitet werden, um fundierte und nachvollziehbare Prüfungen durch Dritte zu ermöglichen.

Abschnitt 3: Potenzielle Maßnahmen und Alternativen

Dieser Abschnitt befasst sich mit den Vor- und Nachteilen möglicher Interventionen und neuer Ansätze, die darauf abzielen, algorithmische Ranking-Praktiken der Social-Media-Plattformen zu reformieren. Dabei werden verschiedene Alternativen beleuchtet – von Ansätzen für mehr individuelle Selbstbestimmung bis zur Neuorientierung der Algorithmen – mit dem Ziel, demokratische, transparente und sichere Online-Räume zu fördern.

3.1 »Erstelle Deinen eigenen Feed«

Der Ansatz der verstärkten Selbstbestimmung der Nutzer:innen (*user agency*), vor allem über einen sogenannten Middleware-Markt, konzentriert sich auf die Schaffung individueller Entscheidungsmöglichkeiten und wettbewerbstauglicher Wege zur Abschwächung negativer Auswirkungen, die mit einer Monopolstellung von Plattformen verbunden sind. In akademischen Debatten wird seit längerem darüber diskutiert, inwiefern den Nutzer:innen eine aktivere Beteiligung an Ranking-Systemen bzw. eine eigene Entscheidung darüber ermöglicht werden könnte, ob oder in welchem Rahmen sie algorithmische Systeme nutzen möchten.

Dezentraler Middleware-Markt als Lösungsansatz

2020 schlug die Working Group on Platform Scale an der Stanford University[II] die Einführung sogenannter Middleware vor, mit der die zentralisierte Machtstellung von Plattformen eingedämmt werden sollte. Den Begriff Middleware definiert die Arbeitsgruppe als »Software und Dienstleistungen, die eine redaktionelle Ebene zwischen den marktbeherrschenden Internetplattformen und den Internetnutzer:innen schaffen«. **Durch einen Einsatz von Middleware soll die konzentrierte Macht der Tech-Unternehmen zur Kontrolle der Informationsflüsse auf ihren Plattformen eingeschränkt und Auswirkungen der plattform-eigenen Algorithmen verringert werden.**

Einen ähnlichen Vorschlag unterbreitet Mike Masnick in seinem Aufsatz »Protocols, not platforms«, in dem er für eine Rückkehr zu dezentralisierten Internetprotokollen, also Regeln für den Austausch von Daten zwischen Endgeräten in einem Rechnernetzwerk, plädiert.[41] Auch dieser Vorschlag zielt darauf ab, die Macht auf individuelle Nutzer:innen oder Dienstleistungsanbieter zu verteilen, statt sie auf wenige Plattformen zu konzentrieren.

Diese Idee ist nicht neu. So bietet Gobo.social den Nutzer:innen mehr Kontrolle über ihre Feeds und die Möglichkeit, bis zu drei Accounts von verschiedenen Plattformen auf einer einzigen Seite zu integrieren, um einen kombinierten Feed anzuzeigen.[42] Ein weiteres Beispiel ist die *Developer Toolbox* von Twitter. Diese beinhaltet insgesamt elf verschiedene Werkzeuge für die individuelle Ausgestaltung von Plattformfunktionen in den Kategorien Kommunikation (*expression tools*), Sicherheit (*safety tools*) und Analyse (*measurement tools*), aus denen die Nutzer:innen wählen können. Mit dem Tool *Block Party* zum Beispiel können Nutzer:innen ihren Online-Alltag möglichst belästigungsfrei gestalten und sich vor Trollen schützen, indem sie unerwünschte Erwähnungen mithilfe von Listen blockieren, bestimmte Nutzer:innen automatisch stummschalten oder sich Unterstützung in der Community holen.[43] Alle Tools werden auf der Grundlage der Qualitäts- und Sicherheitsstandards von Twitter zugelassen, wobei die Anbieter ihre Vermarktungsstrukturen und Preise selbst bestimmen.[44]

Auch die *Perspective API* von Google kann als Middleware-Technologie fungieren und nutzt dabei eine ML-Engine um »toxische« Kommentare zu identifizieren und die Moderation von Unterhaltungen zu verbessern. Die API kann schwere Verstöße, Beleidigungen, Anstößigkeiten, Angriffe auf die Persönlichkeit, Drohungen und sexuell anzügliche Inhalte anhand eines Punktesystems bewerten.[45]

Middleware stellt also einen Lösungsansatz dar, mit dem einzelnen Nutzer:innen über eine Vielfalt an Anbietern mehr Freiheit bei deren Konsumentscheidungen verschafft werden soll. Daphne Keller

[II] Das *Program on Democracy and the Internet* hat im Januar 2020 eine Arbeitsgruppe aufgestellt, die sich mit der Größe, der Reichweite und der Macht digitaler Plattformen befasste. Sie sollte mögliche negative Auswirkungen untersuchen und gegebenenfalls Empfehlungen für Abhilfemaßnahmen aussprechen. Zu der interdisziplinären Gruppe von Wissenschafter:innen gehörten Francis Fukuyama, Barak Richman, Ashish Goel, Roberta R. Katz, A. Douglas Melamed und Marietje Schaake.

vom Stanford's Cyber Policy Center nennt Beispiele für eine mögliche zukünftige Verwendung von Middleware-Software. Nutzer:innen könnten sich etwa für einen Filter von einer der ‚Black Lives Matter'-Bewegung angegliederten Gruppe auf YouTube entscheiden, um speziell rassistischen Inhalte auszublenden. Als weitere Ideen nennt Keller die Nutzung von Faktencheck-Services für Google News, oder eine jugendfreie Version von Facebook, die von Disney angeboten wird.«[46]

Ein Middleware-Markt für algorithmische Ranking-Systeme bringt jedoch auch neue Herausforderungen mit sich.

- Wenn einzelne Nutzer:innen ihr Informationsumfeld selbst gestalten können, könnte dies bewirken, dass bestimmte Gruppen eher Inhalte konsumieren, die ihre eigenen Ansichten, Überzeugungen und Ängste reflektieren. Somit würde sich auch das Denken in Eigen- und Fremdgruppen, also die Einteilung in »wir« und »die anderen«, verschärfen. Nutzer:innen, die sich für eine »positiv filternde« Middleware entscheiden, könnten sich gegen alle negativen Nachrichten abschirmen. Das Umschalten auf einen solchen »Positivmodus« könnte dazu führen, dass sich Nutzer:innen weniger mit gesellschaftlich relevanten Themen und differenzierten Perspektiven auseinandersetzen. So ist es beispielsweise vorstellbar, dass sich das Publikum für Aktivist:innen für soziale Gerechtigkeit verkleinert, wenn Middleware-Technologien zur Abschirmung bestimmter Nutzer:innen vor »negativen« Themen eingesetzt werden.[47] Andere Nutzer:innen dagegen würden sich womöglich bewusst für die Anzeige und die Interaktion mit grenzwertigen oder schädlichen Inhalten entscheiden.

- Ein Middleware-Markt könnte von den verschiedensten Akteuren, einschließlich politischen Parteien, extremistischen Gruppierungen oder feindlichen Staaten, genutzt werden, um ihren jeweiligen Abonnent:innen bevorzugt eigene Inhalte zu präsentieren. Zudem besteht das Risiko, dass bestimmte Nutzer:innen, die besonders anfällig für schädliche Inhalte sind, nicht ausreichend informiert oder aufgeklärt sind, um die Möglichkeit einer Anwendung von Middleware zu erwägen sowie deren mögliche negative Auswirkungen zu überblicken.

- Neben diesen Bedenken wirft ein möglicher Einsatz von Middleware auch datenschutzrechtliche Fragen auf. Anbieter von alternativen Ranking-Systemen müssten sowohl auf öffentlich zugängliche als auch auf plattformeigene Inhalte zugreifen können. So stellt sich beispielsweise die Frage, inwieweit privat geteilte Inhalte von befreundeten Accounts in die Middleware-Technologie einbezogen werden sollten, um einen für das ordnungsgemäße Funktionieren der Middleware ausreichend großen Datenbestand zu schaffen. Ein Einsatz von Middleware zur Kennzeichnung von Inhalten ließe sich als Browser-Erweiterung konzipieren (was ebenfalls gewisse Datenschutzbedenken mit sich bringt). Hingegen würde die Anwendung alternativer Ranking-Algorithmen einen stärker integrierten Rahmen für den Datenaustausch zwischen den Plattformen und Middleware-Anbietern voraussetzen.[48] Die OECD veröffentlichte 2021 einen Bericht über die Rolle, die Maßnahmen zur Datenportabilität und Interoperabilität bei der Förderung des Wettbewerbs – sowohl innerhalb von als auch zwischen Online-Plattformen – spielen können. In dem Bericht werden Standards vorgestellt, die neben einer gemeinsamen Nutzung von Daten in Echtzeit über verschiedene Dienste hinweg (z. B. das *Cross-Posting* von Inhalten auf mehreren Social-Media-Plattformen) auch Kombinationen von Funktionen ermöglichen (z. B. die Anmeldung über ein einziges Konto bei mehreren verschiedenen Online-Diensten).[49]

- Durchsetzung von Middleware erfordert auch, dass sich die Anbieter von Plattformen und Middleware-Anwendungen mit der Aufteilung der Haftung und der ihnen obliegenden Verpflichtungen in Bezug auf Sorgfalts- und Verantwortungspflichten auseinandersetzen. Darüber hinaus wäre ein nachhaltiges Geschäftsmodell für Middleware-Anbieter erforderlich, um ein ausreichendes Angebot und gesunden Wettbewerb zu gewährleisten. Vorstellbar wäre beispielsweise, dass Anbieter von Middleware kostenpflichtige Abonnements anbieten oder dass Plattformen und Middleware-Anbieter sich auf eine Aufteilung der Gewinne einigen.[50] **Fraglich bliebe dabei allerdings, inwieweit alternative Anbieter von Rankings dann tatsächlich andere Metriken verwenden und zur Verbesserung der Qualität des Online-Diskurses beitragen würden.**

Aktive Selbstbestimmung auf Design-Ebene

Etablierte Ansätze der Plattform-Regulierung zielen auf eine Stärkung der Selbstbestimmung der Nutzer:innen auf Design-Ebene ab. In Großbritannien untersuchte das Behavioural Insights Team (BIT) 2021 in Zusammenarbeit mit dem Centre for Data Ethics and Innovation (CDEI) und die Denkfabrik Doteveryone, wie »aktive« Entscheidungsmöglichkeiten für Nutzer:innen geschaffen werden können. Dem Projekt lag der Gedanke zugrunde, dass die Entwicklung von Systemen, die die Nutzer:innen gemäß ihren jeweiligen Präferenzen nutzen können, ein wichtiger Schritt auf dem Weg zu einer »positiven Technologielandschaft« ist. Die Studie definiert aktive Entscheidungen als »Entscheidungen, die die Wünsche der Nutzer:innen unbehindert widerspiegeln und auf einem Verständnis der voraussichtlichen Konsequenzen beruhen«.[51]

Das BIT hat Experimente mit Prototypen für Smartphone-Betriebssysteme, Webbrowser und soziale Medien durchgeführt, die zeigen, wie aktive Entscheidungsmöglichkeiten in drei Online-Kontexten integriert werden könnten. Hierzu führte das BIT drei Online-Experimente mit jeweils etwa 2.000 Teilnehmer:innen durch und testete, wie die alternativen Schnittstellen im Vergleich zu den Kontrollsystemen abschnitten. Das Social-Media-Experiment berücksichtigte gängige Präferenz-Einstellungen wie die Art der Organisation von Inhalten im Feed (chronologisch oder algorithmisch), das Filtern von nicht vertrauenswürdigen Quellen und Datenschutzeinstellungen. Bei dieser Studie kamen drei Versuchsvarianten zum Einsatz, um die Fähigkeit der Nutzer:innen zu testen, bewusste Entscheidungen in Bezug auf ihre Einstellungen zu treffen. Das Experiment für soziale Medien (veranschaulicht in Abbildung 3) umfasste

- einen Slider-Modus, der die manuelle Anpassung innerhalb eines bestimmten Spektrums ermöglicht,
- einen Privat-Modus, der die Entscheidungen in einer einfachen Wahlentscheidung bündelt,
- responsive Schaltflächen, die Entscheidungsmöglichkeiten nach Themen kombinieren, aber nicht in einer einzigen Schaltfläche gebündelt sind.

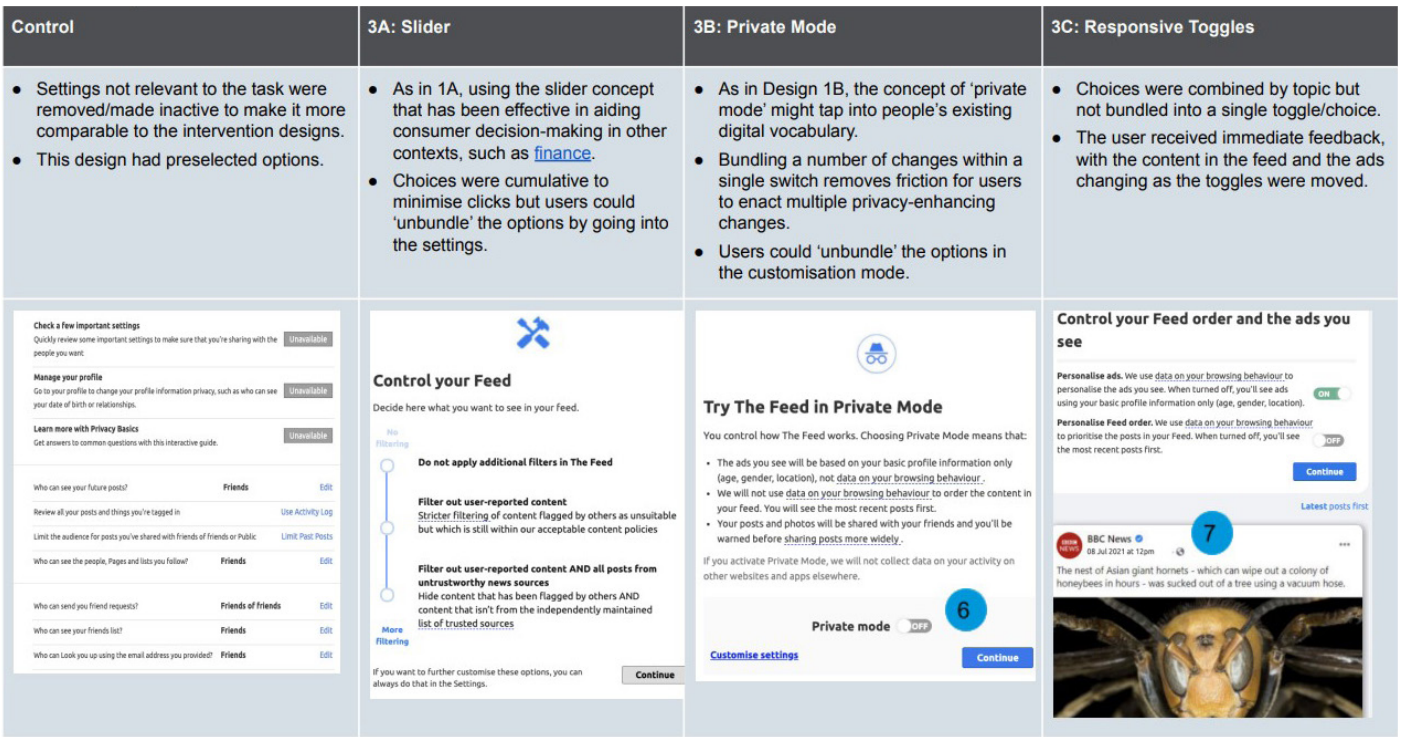

Abbildung 3: Konzeption des Experiments für soziale Medien. Grafik übernommen vom BIT.

Die Ergebnisse der Experimente liefern Erkenntnisse, die für Maßnahmen wie die Einführung von Middleware-Optionen nutzbar gemacht werden können. Dabei fiel zunächst auf, dass die selbst erfassten Metriken offenbar kein guter Indikator für die Fähigkeit der Menschen sind, Entscheidungen im Einklang mit ihren Präferenzen zu treffen. Das zeigt sich beispielsweise darin, dass das selbst angegebene subjektive Gefühl der Kontrolle sich auch dann verbesserte, wenn die Fähigkeit, die Einstellungen präzise nach fiktiven Präferenzen vorzunehmen, und das Verständnis für die Folgen von Entscheidungen sich nicht verbesserten.

Zweitens wurde deutlich, dass die Ausgestaltung der Systeme sorgfältig auf den Wissensstand der jeweiligen Nutzer:innen zugeschnitten sein müssen. Das Feedback deutete darauf hin, dass der Prototyp für das Social-Media-Experiment zu komplex ausgestaltet war. In diesem Zusammenhang wären eine klare Kennzeichnung der Optionen, Transparenz und Tests von zentraler Bedeutung, damit Vereinfachungen und Bündelungen nicht zu einer Benachteiligung bestimmter Nutzergruppen führen und die Nutzer:innen in der Lage sind, ihre Entscheidungen vollständig auf ihre Bedürfnisse abzustimmen.

Schließlich gab es Hinweise darauf, dass die Leistung des Prototyps einer vertrauenswürdigen Drittpartei, bei dem die Teilnehmer:innen Entscheidungen an eine Drittorganisation (die ein Middleware-Anbieter sein könnte) delegieren konnten, von dem Bekanntheitsgrad und der Verbindung dieser Organisation mit der digitalen Welt abhängt. So entscheiden Nutzer:innen sich unter Umständen auch dann für ein bekanntes Tech-Unternehmen, wenn die angebotenen gebündelten Einstellungsmöglichkeiten nicht ihren Präferenzen entsprechen. Zusammenfassend lässt sich sagen, dass es keine »Allzwecklösung« für die Stärkung der Selbstbestimmung der Nutzer:innen gibt, da die Leistungsfähigkeit und Ergebnisse wahrscheinlich vom individuellen Kontext, dem Wissensstand der betreffenden Nutzer:innen und ihrer Vertrautheit mit den einzelnen Anbietern abhängen.

Insgesamt betrachtet könnten Nutzer:innen durch mehr Selbstbestimmung und ein dezentralisiertes Marktumfeld mehr individuelle Entscheidungsmöglichkeiten und Macht erhalten. Theoretisch könnte dieser Ansatz einen gesünderen Wettbewerb unter den Anbietern fördern und damit den Einfluss einiger weniger besonders mächtiger Unternehmen eindämmen. Die Anbieter von Middleware müssten sich jedoch nach wie vor mit den bekannten Herausforderungen des Datenschutzes, der Inhaltsmoderation und nachhaltigen Geschäftsmodellen auseinandersetzen. Hinzu kämen Risiken einer potenziellen Fragmentierung, die zur Zunahme von schädlichen oder grenzwertigen Inhalten und Verhaltensweisen in bestimmten Nutzergruppen führen könnte. So könnte eine individualisierte Auswahl des Informationsumfelds zu einer weiteren Spaltung der Nutzergruppen führen und Akteur:innen, die bewusst negative oder spalterische Inhalte unter ausgewählten Zielgruppen verbreiten, neue Türen öffnen.

Policy Initiative: »Safety by Design«-Rahmenwerk der australischen Regierung

Im Juni 2018 gab die unabhängige Aufsichts- und Aufklärungsbehörde für Online-Sicherheit in Australien, eSafety, die Absicht bekannt, ein *Safety by Design (SbD)*-Rahmenwerk und dazugehörige Prinzipien zu entwickeln. Dieses Rahmenwerk ist ein umfassendes und iterativ angelegtes Programm, das Tech-Unternehmen dabei unterstützen soll, die Rechte der Nutzer:innen und deren Sicherheit in das Design und die Funktionalität ihrer Produkte und Dienstleistungen zu integrieren. Die *SbD*-Prinzipien fußen auf den Menschenrechten und wurden auf Grundlage von Informationen entwickelt, die im Rahmen der Forschungs- und Berichtsprogramme von eSafety, von Sensibilisierungsprogrammen sowie bei Konsultationen mit Industrie und wichtigen Interessengruppen gesammelt wurden.

Die zweite Phase des *SbD*-Rahmenwerks konzentriert sich auf die Entwicklung eines Leitfadens, der den Industriepartnern bei der Umsetzung der Prinzipien und der Verbesserung der Online-Sicherheitspraktiken und -Maßnahmen helfen soll. Diese Orientierungshilfen sollen von Social-Media-Plattformen als auch von verschiedenen Online-Segmenten in den Bereichen Gaming, Dating und Banking verwendet werden.

Das Rahmenwerk beinhaltet Prinzipien zur Stärkung der Selbstbestimmung und Autonomie der Nutzer:innen. Dafür sollten Online-Dienste technische Maßnahmen und Tools bereitstellen, mit denen Nutzer:innen ihre Online-Sicherheit selbst beeinflussen können und die standardmäßig auf die sichersten Datenschutz- und Sicherheitseinstellungen konfiguriert sind. Weiterhin sollten sie »den Einsatz technischer Funktionen nutzen, um Risiken und Schäden abzumildern«. Dies bezieht sich auf technische Mechanismen, mit denen Inhalte gefiltert werden können – was nicht zwangsläufig bedeute, diese zu entfernen, sondern beispielsweise zu Altersbeschränkungen oder zwischengeschalteten Warnseiten führen könnte. Schließlich sollten die Online-Dienste ihr Design und ihre Funktionen »bewerten«, um sicherzustellen, dass die Risikofaktoren für alle Nutzer:innen, insbesondere für diejenigen mit Diskriminierungsmerkmalen, vor einer öffentlichen Freigabe der Produkte entschärft werden.

Im Zuge einer fortschreitenden Entwicklung des Online-Ökosystems zielt der »Safety by Design«-Ansatz (ähnlich wie »Privacy by Design« und »Security by Design«-Ansätze) darauf ab, Schäden abzuwehren, bevor sie entstehen. Die Anwendung der Prinzipien seitens der Unternehmen bleibt indes freiwillig. Obgleich es demzufolge auch keine Durchsetzungsmechanismen für deren Einhaltung gibt, sollen sie als Ausgangspunkt für die Selbstregulierung dienen und die durchsetzbaren Koregulierungs- und Regulierungsanforderungen des australischen *Online Safety Act 2021* untermauern. Das Gesetz sieht vor, dass Branchenverbände oder ähnliche Gremien »verbindliche Regeln« ausarbeiten, die vom sogenannten eSafety Commissioner anerkannt werden können, sofern sie »angemessene gesellschaftliche Sicherheitsanforderungen« erfüllen. Falls dies nicht gelingen sollte, kann eSafety selbst einen »Standard« festlegen.[52] Das Gesetz legt grundlegende Anforderungen (*Basic Online Safety Expectation*) für Anbieter von Online-Diensten fest, um Nutzer:innen proaktiv vor missbräuchlichem Verhalten und schädlichen Inhalten zu schützen. eSafety kann zudem von den Anbietern einen Bericht darüber einfordern, wie sie einzelne oder sämtliche dieser Anforderungen erfüllen.[53]

3.2 Qualitätsorientierte Ranking-Algorithmen

Neue Vorschläge zielen auf ein Umdenken der Unternehmen in Bezug auf die dem algorithmischen Ranking zugrunde gelegten Metriken ab, damit künftig andere Arten von Inhalten und Verhaltensweisen belohnt werden. So stellte Renee DiResta vom Stanford Internet Observatory fest: »Plattformen können sich dafür entscheiden, fragwürdige Inhalte wie bei *Pizzagate*[III] auf ihrer Seite zuzulassen, und gleichzeitig dafür sorgen, dass diese Inhalte nicht algorithmisch verstärkt oder den Nutzer:innen proaktiv empfohlen werden«.[54] Das Überdenken der Metriken zielt demnach nicht darauf ab, schädliche Inhalte vollständig zu entfernen. Vielmehr sollten die Algorithmen so umgestaltet werden, dass sie einen weniger spalterischen Diskurs fördern. **Die Idee eines »qualitätsorientierten« Rankings ist nicht neu. Was die tatsächliche Umsetzung seitens der Unternehmen betrifft, konnte sich der Ansatz bisher zwar nicht erfolgreich durchsetzen. Dennoch lohnt es sich, eine Umsetzung weiter zu forcieren und dabei gegebenenfalls auch regulatorische Rahmenbedingungen zu diskutieren.**

Industriepraxis: Facebooks »Remove, Reduce, Inform«-Strategie

Seit 2016 setzt Facebook bei seinem News Feed zur Bekämpfung irreführender oder schädlicher Inhalte auf eine Strategie des Entfernens, Reduzierens und Informierens (*Remove, Reduce, Inform*). **Das bedeutet, dass der News Feed nicht nur Inhalte entfernt, die gegen die Gemeinschaftsstandards verstoßen, sondern auch »qualitativ minderwertige« Inhalte wie Engagement Baiting (d. h. Taktiken, bei denen Personen dazu gedrängt werden, mit Facebook-Beiträgen zu interagieren, um künstlich größere Reichweite zu generieren) oder Websites mit wenig Substanz sowie störende Werbeinhalte zurückstuft.** Dabei verwendet der Algorithmus für den News Feed Signale, die beispielsweise bestimmen, ob der Facebook-Traffic einer Domain in hohem Maße unverhältnismäßig zu ihrer Platzierung im Webgraphen ist.[55] Damit ist gemeint, dass Seiten und Hyperlinks im Internet als Knoten (*nodes*) und Bögen (*arcs*) in einem gerichteten Graphen betrachtet werden können. Die Teams von Facebook überprüften Websites, die von und zu Facebook verlinkt sind, um solche zu identifizieren, die »wenig substanzielle« Inhalte und »eine große Anzahl von verstörenden, schockierenden oder bösartigen Anzeigen« enthielten. Anschließend wurden Algorithmen verwendet, um neue Websites mit qualitativ minderwertigen Inhalten zu identifizieren. Gemäß den Richtlinien für die Verbreitung von Inhalten von Meta (Stand: Oktober 2022)[56] stuft der News Feed ebenfalls folgende Inhalte zurück:

- Kommentare von niedriger Qualität (d. h. Kommentare, die keinen sinnvollen Beitrag zum Diskurs über einen Beitrag leisten);
- Events bzw. Veranstaltungen von niedriger Qualität (insbesondere Events, für die keine Angaben zum Zeitpunkt, Ort und/oder Anmeldeinformationen vorliegen);
- Videos von geringer Qualität (insbesondere Videos, die als statisch oder animiert eingestuft werden bzw. die als Dauerschleifen, reine Umfragen oder voraufgezeichnete Videos konzipiert sind);
- Seiten, die Videos aus Fremdquellen veröffentlichen (insbesondere Videos, die anderen Quellen entstammen und mit geringem Mehrwert wiederverwendet werden);
- Seiten, bei denen es sich vermutlich um Spam handelt;
- Reißerische Inhalte zum Thema Gesundheit und kommerzielle Beiträge zu Gesundheitsthemen;
- Domains mit wenig Originalinhalten (z. B. mit einer großen Menge an Inhalten fremder Urheber);
- Falschmeldungen, die durch einen Faktencheck entlarvt wurden;
- Unauthentisches Teilen (insbesondere Seiten, deren Verhalten die Zahl der Aufrufe oder das Engagement künstlich erhöht);
- Links zu Domänen und Seiten mit hoher »Click-Gap« (d. h. Domänen, die einen unverhältnismäßig hohen Anteil ihres Traffics direkt von Facebook erhalten, verglichen mit der Menge des Traffics aus dem übrigen Internet);

[III] Einer als »Pizzagate« bekannten viralen Verschwörungstheorie zufolge haben Hillary Clinton und demokratische Abgeordnete, die Essensbestellungen bei einer Pizzeria namens Comet Ping Pong im US-amerikanischen Washington aufgaben, angeblich in Wirklichkeit einen Code verwendet, um über minderjährige Prostituierte zu sprechen.

- Beiträge von Nachrichtenanbietern, die weithin als nicht vertrauenswürdig gelten;
- Beiträge von Seiten, die ihre Verbreitung künstlich in die Höhe treiben;
- Beiträge von Personen, die Hyper-Sharing in Gruppen verwenden;
- Nachrichtenartikel aus Fremdquellen.

Darüber hinaus plant Facebook, dass der News Feed den Nutzer:innen kontextbezogene Informationen zu Beiträgen bereitstellt, die »Glaubwürdigkeitssignale« (*credibility signals*) enthalten. So könnten beispielsweise Informationen über die Seite angezeigt werden, die den Originalartikel veröffentlicht hat. Ebenso könnte Kontext von externen Expert:innen oder Drittorganisationen (Wikipedia) bereitgestellt oder auf »ähnliche Beiträge oder weitere Inhalte vom Herausgeber« verwiesen werden.[57] **Trotz dieser Bemühungen hat Facebook noch keine klaren Definitionen für Begriffe wie »Herabstufung« (auch »downranking« genannt) vorgelegt. Unklar ist somit unter anderem, wie lange Nutzer:innen scrollen müssen, um herabgestufte Inhalte zu sehen. Eine Bewertung des Erfolgs dieser Strategie ist daher schwierig. Außerdem bleibt unklar, ob minderwertige Inhalte wie Engagement Baiting, Spam oder Seiten, die Videos aus Fremdquellen posten, in gleichem Maße herabgestuft werden wie sensationsheischende Gesundheitsinformationen oder per Faktencheck entlarvte Fehlinformationen.**[58]

Industriepraxis: »Search Quality Rating« von Google

Googles sogenanntes *Search Quality Rating* dient als Beispiel für die potenzielle Weiterentwicklung von Metriken, um »Qualität« (etwa die Verlässlichkeit von Quellen) in die Gewichtung algorithmischer ML-Modelle aufzunehmen. Unterdessen kritisieren Forscher:innen die Ergebnisse der Suchalgorithmen, da sie in der Vergangenheit immer wieder schädliche Inhalte, einschließlich diskriminierender und irreführender Inhalte, nicht ausreichend zurückgestuft oder entfernt haben. Insofern sollten die Suchergebnisse von Google unbedingt weiterhin auf ihren Anspruch getestet werden, »Qualität« zu gewährleisten.

Googles Suchalgorithmen gewichten – neben der Bedeutung der Suchanfrage, der Relevanz der Inhalte in Bezug auf die jeweilige Anfrage, Kontext und Nutzerfreundlichkeit – auch die Qualität der Inhalte auf einer Website. In einem im Februar 2019 veröffentlichten White Paper stellt Google explizit fest, dass Ranking-Algorithmen ein wichtiges Instrument im Kampf gegen Desinformation seien.[59] Demnach würden die Suchalgorithmen minderwertige oder schädliche Ergebnisse auf »weniger sichtbare Positionen« in den Suchergebnissen zurückstufen. In dem Dokument wird darauf hingewiesen, dass die Algorithmen weder feststellen können, ob ein Inhalt zu aktuellen Ereignissen wahr oder falsch ist, noch könnten sie die Absicht des Urhebers allein anhand des Seiteninhalts einschätzen. Sie sind jedoch in der Lage, Manipulations- oder Täuschungstaktiken wie Spam-Verhalten zu identifizieren.[60] **Unterdessen zielen die Ranking-Algorithmen darauf ab, »hochwertige Seiten« (*High Quality Pages*) nach Google-Richtlinien (Search Quality Rater Guidelines) zu priorisieren. Demnach werden folgende Merkmale evaluiert:**

- Ein hohes Maß an Sachkompetenz (*Expertise*), Maßgeblichkeit (*Authoritativeness*) und Vertrauenswürdigkeit *(Trustworthiness*) – kurz: E-A-T.
- Eine hinreichende Menge an hochwertigem Hauptinhalt (*Main Content, MC*), einschließlich eines beschreibenden oder hilfreichen Titels. Qualitativ hochwertiger MC muss sachlich korrekt sein und von einem wissenschaftlichen Konsens gestützt werden, sofern ein solcher Konsens besteht.
- Zufriedenstellende Informationen über die Website und/oder Informationen darüber, wer für die Website verantwortlich ist – sofern die Seite hauptsächlich zum Online-Shopping dient oder Finanztransaktionen beinhaltet, sollte sie zufriedenstellende Informationen über den Kundenservice enthalten.
- Positive Bewertung einer Website, die für den MC auf der Seite verantwortlich ist (oder positive Bewertung des Erstellers des MC, sofern dieser nicht mit dem Ersteller der Website identisch ist).[61]

2017 kündigte Google zudem eine Partnerschaft mit *The Trust Project*[IIII] an, um Nachrichtenseiten bei der Verwendung von acht Indikatoren für Vertrauenswürdigkeit (Trust Indicators) zu helfen[62]:

1. Bewährte Praxis (*Best Practices*): Was sind die Standards der Nachrichtenredaktion? Wer finanziert sie? Wie lautet ihre Mission?
2. Fachwissen: Wer hat den Bericht verfasst? Nähere Angaben über die Journalist:innen, einschließlich ihrer Fachkompetenzen?
3. Kennzeichnung: Um welche Art von Inhalt handelt es sich? Gibt es Kennzeichnungen, die dazu dienen, Meinungen, Analysen und gesponserte Artikel von Nachrichten zu unterscheiden?
4. Quellenangaben: Was ist die Quelle? Gibt es Zugang zu Quellen, die hinter Fakten und Behauptungen investigativer oder ausführlicher Berichte stehen?
5. Methoden: Warum ist der Inhalt entstanden? Gibt es Informationen darüber, warum sich die Journalist:innen entschieden haben, über ein Thema zu berichten, und wie sie dabei vorgegangen sind?
6. Vor Ort gesammelte Informationen: Erfolgte die Recherche direkt vor Ort, mit fundierten Kenntnissen der lokalen Situation oder der örtlichen Gemeinschaft?
7. Vielfältige Ansichten: Inwiefern bemüht sich die Redaktion, verschiedene Perspektiven einzubringen?
8. Möglichkeiten für Feedback: Können die Leser:innen partizipieren? Inwiefern bemüht sich die Redaktion, die Öffentlichkeit bei der Festlegung von Schwerpunkten der Berichterstattung einzubeziehen?

Google hat bereits erläutert, dass diese auf E-A-T basierenden Kriterien und die Bewertungsdaten (*Rater Data*) nicht *direkt* von den Ranking-Algorithmen verwendet werden. Vielmehr werden sie als Feedback verwendet, damit Ranking-Teams nachvollziehen können, ob die Algorithmen funktionieren.[63]

So sollen diese Kriterien auch in dem wohl bekanntesten Signal für Maßgeblichkeit (*Authoritativeness*), dem sogenannte PageRank, dargestellt werden. Der PageRank berücksichtigt dabei Anzahl der Links im Web, Link-Attribute, Ankertext und die Wahrscheinlichkeit, angeklickt zu werden. Vereinfacht bedeutet dies: Je mehr Quellen auf eine Seite verlinken, desto wertvoller sind die Informationen auf dieser Seite und desto wahrscheinlicher ist es, dass die Nutzer:innen sie besuchen. 2006 hat Google die Algorithmen so umgestaltet, dass einige wenige vertrauenswürdige Quellen, die sogenannten *Seed*-Seiten, ausgewählt werden und die Qualität anderer Seiten anhand der von diesen Seiten ausgehenden Links bewertet wird. So stuft Google beispielsweise die New York Times als *Seed*-Seite ein, da sie ein breites Spektrum an Themen abdeckt, die viele Nutzer:innen interessieren, und viele ausgehende Links enthält.[64]

Das Integrity Institute schlägt vor, dass die algorithmischen Ranking-Systeme der Social-Media-Plattformen Kriterien wie den PageRank einbeziehen können, um Seiten oder Accounts zu überwachen und ihnen nach Überprüfung der Medienkompetenz eine Punktzahl (*score*) zuzuweisen.[65] Auf diese Weise könnten Inhalte aus anonymen Quellen (ausgenommen allerdings solche, die Inhalte aus legitimen Gründen anonym veröffentlichen, wie beispielsweise Whistleblower:innen oder Menschenrechtsaktivist:innen), Inhalte aus Quellen, die systematisch Inhalte von anderen Urhebern kopieren, Inhalte aus Quellen, die vernetzte Programme (Bots) nutzen, oder Inhalte, die gegen die Gemeinschaftsstandards verstoßen, bei einer Überprüfung der Medienkompetenz durchfallen. **Wichtig bei diesem Ansatz bleibt, dass die Entwicklung und Verwendung von »Qualitätskriterien« als Metriken sowie die Methoden, mit denen sie getestet und durchgesetzt würden, mit transparenten Stakeholder-Konsultationen und im Dialog mit der Zivilgesellschaft und externen Forscher:innen verknüpft werden sollten.**

[IIII] Das *The Trust Project* wurde von der Journalistin Sally Lehrman ins Leben gerufen und wird u.a. von Craig Newmark Philanthropies, Google, der John S. and James L. Knight Foundation und dem Democracy Fund gefördert.

3.3 Positive Friktion, Nudges und brückenbildendes Ranking

Ein Aspekt des Problems der Ranking-Algorithmen ist die Leichtigkeit und Geschwindigkeit, mit der Nutzer:innen Inhalte ununterbrochen veröffentlichen und teilen können, ohne auf nennenswerte Reibungen oder Unterbrechungen zu stoßen. Sogenannte »positive Friktion« zielt daher darauf ab, das Veröffentlichen und die Interaktionen zu entschleunigen, um den Nutzer:innen die Möglichkeit zu geben, ausreichend zu reflektieren, bevor sie Inhalte teilen.[66] Eine solche Zielsetzung könnte in algorithmische Ranking-Praktiken integriert werden.

Facebook-Whistleblowerin Frances Haugen hatte in der Vergangenheit betont, dass sich viele Superuser insbesondere am späten Abend mit grenzwertigen Inhalten beschäftigen. Eine Entschleunigung des algorithmischen Rankings in den Abendstunden könnte daher dazu beitragen, dass Superuser früher abschalten und Algorithmen weniger toxische Signale von solchen Nutzer:innen erhalten. Um den unverhältnismäßigen Einfluss von Superuser-Aktivitäten zu verringern, sprach sich Haugen für eine Art Feuermelder-Prinzip (*break glass measures*) aus und meinte damit konkret Maßnahmen, mit denen die Anzahl der Shares eines jeden beliebigen Inhalts einfach begrenzt werden könnte. Eine Entschleunigung im Informationsfluss entstünde auch durch das Hinzufügen einer weiteren Entscheidungsebene, da die Nutzer:innen nun gefordert wären, Inhalte zu kopieren und einzufügen, wenn sie diese teilen möchten.[67]

Ein vom MIT Sloan veröffentlichtes Arbeitspapier untersuchte, wie und warum sich Fehlinformationen über COVID-19 in den sozialen Netzwerken verbreiten, und testete, ob ein einfacher technischer Eingriff (auch Nudge genannt) diese Verbreitung begrenzen könnte.[68] In Experimenten beobachteten die Forscher:innen, dass Nutzer:innen, die zum Nachdenken über den Wahrheitsgehalt eines Inhalts angeregt wurden, beim Teilen von wahren oder falschen Nachrichten kritischer wurden. **Dabei fiel besonders auf, dass Nutzer:innen, die zuerst den Wahrheitsgehalt einer Nachricht bewerteten, falsche Nachrichten weniger häufig teilten als wahre. Darüber hinaus stellen die Forscher:innen fest, dass die kumulativen Auswirkungen einer solchen Maßnahme möglicherweise wesentlich größer sind, als die Untersuchung der getesteten Einzelpersonen zeigt.**

Policy Initiative: Gesetzesentwurf für den Social Media NUDGE Act in den USA

Im Februar 2022 haben die Senatorinnen Amy Klobuchar (D-MN) und Cynthia Lummis (R-WY) den *Nudging Users to Drive Good Experiences on Social Media Act* (kurz: *Social Media NUDGE Act*) im US-Senat eingebracht.[69] Der Gesetzentwurf sieht eine Zusammenarbeit zwischen der National Science Foundation (NSF) und der National Academy of Sciences, Engineering, and Medicine (NASEM) vor, damit diese »kontinuierliche Studien zur Ermittlung inhaltsneutraler Maßnahmen« durchführen, die die großen Social-Media-Plattformen umsetzen könnten, um »die Schäden der algorithmischen Verstärkung und der Social-Media-Sucht zu verringern«.

Der Gesetzentwurf definiert »inhaltsneutrale Maßnahmen« (*content-agnostic interventions*) als Maßnahmen, die – unabhängig von der Art des Inhalts – die Erfahrung einzelner Nutzer:innen verändern. Als Maßnahmen werden beispielsweise »Warnungen zur Bildschirmzeit«, »Kennzeichnungen und Warnungen, die Nutzer:innen dazu auffordern, nutzergenerierte Inhalte vor Weiterverbreitung zu lesen oder zu überprüfen« oder »Aufforderungen an die Nutzer:innen, die ihnen helfen können, manipulative und gezielte Werbung zu erkennen« aufgelistet. Die US-amerikanische Federal Trade Commission (FTC) müsste ein Regelsetzungsverfahren einleiten, um festzulegen, welche der empfohlenen Maßnahmen verbindlich vorgeschrieben werden sollten.

Der Gesetzentwurf, der auf mehr Transparenz und Meldepflichten abzielt, verpflichtet die Plattformen außerdem dazu, Informationen über die Einhaltung der Vorschriften, die Wirkung der Maßnahmen und Statistiken über die geforderten Änderungen und Inhalte auf ihren Plattformen zu veröffentlichen (einschließlich der Gesamtzahl der Aufrufe für jeden öffentlich sichtbaren Inhalt, der im Laufe des Monats veröffentlicht wurde, sowie Stichproben des Inhalts dieser Posts). Da Verstöße als unfaire oder betrügerische Handlungen oder Praktiken gewertet werden, überträgt der Gesetzentwurf die Durchsetzungskompetenz an die FTC.

Der Gesetzesentwurf unterscheidet sich insofern von verschiedenen anderen Gesetzen, die dem Kongress vorgelegt wurden und die darauf abzielen, Social-Media-Plattformen zu regulieren, als er sich nicht auf bestimmte Arten von Inhalten konzentriert, sondern auf die Art und Weise, *wie* sich Inhalte online verbreiten.

Ellen P. Goodman von der Rutgers Law School merkt an, dass es zwar »gut ist, Forschungsinitiativen zu fördern, die sich mit der Frage befasse, welche Arten von Maßnahmen wirksam sind«. Ihr zufolge stünden dem Gesetzesentwurf wegen der verwendeten Begriffe rund um »inhaltsunabhängige« (*content-agnostic*) und »inhaltsneutrale« (*content-neutral*) Maßnahmen jedoch die üblichen Herausforderungen bevor, die alle Gesetzesinitiativen zur Inhaltsmoderation von Plattformen im Zusammenhang mit dem Recht auf freie Meinungsäußerung zu bewältigen hätten. Goodman argumentiert, dass konzeptionelle Eingriffe zur Herabstufung bestimmter Inhalte oder zur Lenkung von Nutzer:innen auf andere Inhalte wahrscheinlich nicht als »inhaltsunabhängig« angesehen werden können, insbesondere weil diese Maßnahmen nicht mehr *per se* inhaltsunabhängig seien, sobald sie mit »gezielter Werbung« oder als »manipulativ« eingestuften Inhalten in Verbindung gebracht werden müssten.[70]

Industriepraxis: Twitters »Birdwatch« (Community Notes)

Um irreführenden Tweets beizukommen, führte Twitter 2021 in den USA im Rahmen des Projekts »Birdwatch« ein Community-Moderationsprogramm ein. Das Pilotprojekt blieb zunächst von Twitter getrennt. Das Grundprinzip des Ansatzes ist, dass Nutzer:innen, die sich bei Birdwatch anmelden, an der Lösung des Problems mitwirken, indem sie Tweets identifizieren, die sie für irreführend halten, sogenannte »Notes«, also Anmerkungen, zum Kontext schreiben und die Qualität der Notes anderer Mitwirkender bewerten.

Bei der Anzeige von Birdwatch Notes geht es statt um Mehrheitsregeln oder Beliebtheit darum, dass sie von Nutzer:innen als hilfreich bewertet werden, die tendenziell anderer Meinung sind.[71] Um in einem Tweet angezeigt zu werden, müssen Notes also von Nutzer:innen als »hilfreich« eingestuft werden, die in ihren früheren Bewertungen eher anderer Meinung waren. Dadurch soll die Wahrscheinlichkeit dafür erhöht werden, dass der den Tweets hinzugefügte Kontext für ein breites Publikum hilfreich ist. Die Algorithmen, die bestimmen, welche Notes letztendlich als »hilfreich« oder »nicht hilfreich« angezeigt werden, basieren auf dem Ansatz der Matrixfaktorisierung, also auf der Grundlage einer Bewertungsmatrix für die Notes.

Die Technik, die nach einem Zusammenhang sucht, mit dem die Affinität bestimmter Nutzer:innen zu bestimmten Inhalten erklärt werden kann, wurde ursprünglich 2006 von Funk im Rahmen des Wettbewerbs Netflix Prize erforscht. Mit dem Preis sollte ein verbesserter Collaborative-Filtering-Algorithmus (also algorithmische Empfehlungen basierend darauf, wie andere Nutzer:innen in der Vergangenheit mit Inhalten interagiert haben) gefunden werden. Damit eine Note einen hohen Intercept-Term (d. h. der Wert ihrer Nützlichkeit) erhält, muss sie von Bewerter:innen mit unterschiedlichen Standpunkten (Faktoreinbettungen) als hilfreich bewertet werden. **Auf diese Weise soll der Algorithmus Notes mit breiter Wirkung über verschiedene Standpunkte hinweg erfassen.**

Birdwatch zeigt die beiden »Erklärungs-Tags« an, die von den meisten Bewerter:innen vergeben wurden, um zu erklären, warum sie die Note als hilfreich oder nicht hilfreich bewertet haben (z. B. für hilfreiche Notes die Tags »UnbiasedLanguage«, »UniqueContext«, »Empathetic«, »GoodSources«, »ImportantContext«). Der Algorithmus von Birdwatch ist einschließlich aller zugrundeliegenden Daten öffentlich auf GitHub verfügbar.[72]

Erste Rückmeldungen deuten darauf hin, dass die Nutzer:innen die Notes als eine »Meinung der Community« (und nicht die von Twitter oder einer zentralen Behörde) wertschätzten. Insbesondere wurde gewürdigt, dass die Notes nützlichen Kontext lieferten, der den Nutzer:innen half, einen Tweet besser zu verstehen und zu bewerten (anstatt sich darauf zu beschränken, Inhalte als »wahr« oder »falsch« zu kennzeichnen).[73] Laut den Umfragen von Twitter mit 3.000 bis 19.000 Teilnehmer:innen, die zwischen August 2021 und August 2022 durchgeführt wurden, ist die Wahrscheinlichkeit dafür, dass eine Person, die eine Birdwatch Note sieht, dem Inhalt eines potenziell irreführenden Tweets zustimmt, im Mittel um 20 bis 40 Prozent geringer als bei Nutzer:innen, die den Tweet ohne zusätzlichen Kontext sehen. Im Oktober 2022 kündigte Twitter an, dass das Unternehmen die Sichtbarkeit von Birdwatch Notes, die von Mitwirkenden mit »hilfreich« bewertet wurden, auf alle Nutzer:innen in den USA ausweiten wird.[74]

Aviv Ovadya beschreibt Twitters Birdwatch als brückenbildendes Ranking (*bridging-based ranking*). Er schlägt vor, Ranking-Systeme so auszugestalten, dass Inhalte belohnt werden, die »zu positiven Interaktionen zwischen verschiedenen Nutzergruppen führen, selbst wenn das Thema an sich potenziell spalterisch ist«.[75] Ein solcher brückenschlagender Algorithmus würde als »zentripetal wirkendes Ranking« fungieren und die Beiträge belohnen, die in Bezug auf die positiven Reaktionen der verschiedenen Zielgruppen am besten abschneiden. Bei der Brückenbildung ginge es nicht darum, gegensätzliche Meinungen zu vertreten (was zu weiteren Meinungsspaltungen führen kann), sondern Inhalte höher zu bewerten, die mehr »gemeinsame Grundlagen« aufweisen. Ovadya merkt an, dass der Ansatz der Brückenbildung weitere Forschung zu dem Thema erfordert, wie Bewertungs- und Rankingmechanismen Signale wie Benutzerinteraktionen nutzen können, um plattformweit »konstruktive Konflikte« von »destruktiven Konflikten« zu unterscheiden. Auch die Frage, ob ein

solches Ranking aufgrund von verzerrten kollektiven Bewertungen zu einer geringeren Qualität führen könnte, rechtfertigt gewisse Bedenken. Schließlich geht es beim »Bridging« nicht um den Inhalt an sich, sondern um die Vielfalt der Nutzer:innen, die positiv mit diesem Inhalt interagieren. Letztendlich ist bei der Entwicklung eines »brückenbildenden« Rankings – wie bei jeder anderen Maßnahme auch – ein ausreichendes Maß an Transparenz erforderlich, um eine Überprüfbarkeit der Ergebnisse zu gewährleisten.

Zusammenfassend ist festzustellen, dass der Ansatz des brückenbildenden Rankings hauptsächlich auf Lösungen für spalterische Inhalte in Online-Diskursen und beim Nutzer:innenverhalten abzielt. Dabei kombiniert der Ansatz verschiedene Elemente mehrerer oben vorgeschlagener Maßnahmen. Er berücksichtigt die Diskussionen über die Anreize hinter den Metriken, was deren Abkehr von einer Belohnung des Engagements hin zur Belohnung einer »Vielfalt des Publikums« betrifft. Gleichzeitig werden die potenziellen Vorteile von Crowdsourcing, Faktenchecks und Anreizen zur Prüfung des Wahrheitsgehalts genutzt. **Ein Vorteil dieses Ansatzes dürfte sein, dass er sich nicht direkt mit der Schädlichkeit, Unwahrheit oder spalterischen Tendenz des Inhalts selbst befasst, sondern darauf abzielt, die Interaktionen der Nutzer:innen einzubeziehen, um einen Kontext zu schaffen, der für Nutzer:innen, die zu Meinungsverschiedenheiten neigen, hilfreich sein könnte. Brückenbildende Algorithmen befinden sich noch in der Frühphase der Entwicklung. Dieser Ansatz zeigt jedoch vielversprechende Perspektiven auf, die von der Forschungsgemeinschaft, den Plattformen selbst und den politischen Entscheidungsträger:innen weiter untersucht werden sollten.**

Fazit

Angesichts der gesellschaftlichen Auswirkungen des Informationskonsums und Nutzungsverhaltens, die durch den Einsatz algorithmischer Ranking-Systeme auf Social-Media-Plattformen beeinflusst werden, ist ein inklusiver, transparenter und demokratischer Ansatz erforderlich, der von einer Vielzahl an Stakeholdern mitgetragen werden muss. Der Schlüssel für faktenbasierte politische Debatten und Maßnahmen liegt in der Transparenz der unternehmensinternen Methodik (welche Ziele verfolgen die Algorithmen, welche Auswirkungen sind erwünscht und werden durch entsprechende Anreize belohnt?) sowie der algorithmischen Experimente zu den (unbeabsichtigten) Auswirkungen von Metriken auf Inhalte und Diskurse.

Algorithmisches Ranking ist weder neutral noch unveränderbar. Im Gegenteil: Entscheidungen der Unternehmensführung beeinflussen die Metriken von Algorithmen kontinuierlich. Diese Metriken können angepasst werden, damit sie zur Stärkung eines demokratischen, transparenten und sicheren Online-Umfelds beitragen.

Angesichts der methodischen und epistemischen Grenzen in der Forschung bleiben »unknown unknowns« und eine Wissenslücke zwischen den Plattformen auf der einen Seite und den politischen Entscheidungsträger:innen, Aufsichtsbehörden sowie Forschungsgemeinde auf der anderen Seite. Liberaldemokratische Staaten sollten in Forschungsprojekte investieren, die darauf abzielen, eine umfassendere Faktengrundlage in Bezug auf die gesellschaftlichen Auswirkungen des algorithmischen Rankings zu schaffen. Um ein besseres Verständnis der algorithmischen Systeme zu ermöglichen, sind transparente und gemeinsame Bemühungen von zentraler Bedeutung sowohl für die Entwicklung erfolgversprechender politischer Maßnahmen als auch für die Bewertung ihrer Wirksamkeit.

Gleichzeitig sollten diese Staaten in die weitere Ausarbeitung bereits bestehender Ansätze investieren, die darauf abzielen, die Risiken eines auf Engagement ausgelegten Rankings zu bekämpfen. **Wissenschaftliche Evidenz für die Wirksamkeit der Änderungen von Ranking-Systemen kann dazu beitragen, vorgeschlagene Interventionen abzustimmen und so zu integrieren, dass sichere und demokratische Online-Diskurse gestärkt werden.**

Damit algorithmische Ranking-Praktiken der Plattformen im Dienste eines transparenten und sicheren Online-Umfelds funktionieren können, sind letztlich Interventionen auf verschiedenen Ebenen erforderlich. **Wenngleich sich einige dieser Maßnahmen als unwirksam oder kontraproduktiv erweisen, könnten andere die gegenwärtigen Praktiken, wie freie und demokratische Gesellschaften mit Ranking-Algorithmen interagieren, grundlegend verändern.** Um eine kontinuierliche Neubewertung und Überprüfung dieser Algorithmen zu gewährleisten, sind ressortübergreifendes und vielseitiges Expertenwissen sowie unterschiedliche Betrachtungsweisen von entscheidender Bedeutung. Insgesamt ist eine Intensivierung der Anstrengungen von Seiten aller Beteiligten dringend erforderlich, um Wissenslücken und Macht-Ungleichgewichte zu überwinden.

Endnoten

1 Koshiyama, A. et al. (2021) Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. Abrufbar unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3778998

2 Burns, E. (2022). Machine Learning. TechTarget. Abrufbar unter: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

3 Eine qualitative Studie des ISD konnte zum Beispiel nachweisen, wie die Algorithmen von YouTube dazu beitragen, frauenfeindliche, antifeministische und extremistische Inhalte an australische Jungen und junge Männer zu vermitteln. Eine Untersuchung zeigt, dass YouTubes Kurzvideo-Feature Shorts Inhalte als Reaktion auf das Verhalten der Nutzer:innen aggressiver zur optimieren scheint und innerhalb eines relativ kurzen Zeitraums Videos mit extremeren Inhalten empfiehlt. Siehe: Thomas, E. & Balint, K. (2022). Algorithms as a Weapon Against Women: How YouTube Lures Boys and Young Men into the 'Manosphere'. *ISD*. Abrufbar unter: https://www.isdglobal.org/isd-publications/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/

4 »Make Instagram Instagram again« (2022). *change.org*. Abrufbar unter: https://www.change.org/p/make-instagram-instagram-again-saveinstagram; Kramer (2022). Tiktokisierung von Instagram: Adam Mosseri rechtfertigt sich nach Kritik von Kylie Jenner. t3n. Abrufbar unter: https://t3n.de/news/instagram-full-screen-ui-kylie-jenner-1488364/

5 Q2 2022 Earnings (2022). Meta Investor Relations. *Meta*. Abrufbar unter: https://investor.fb.com/investor-events/event-details/2022/Q2-2022-Earnings/default.aspx

6 Patel, N. (13. September 2022). Everyone knows what YouTube is – few know how it really works. *The Verge*. Abrufbar unter: https://www.theverge.com/2022/9/13/23349037/mark-bergen-youtube-creators-tiktok-algorithm

7 Eine Untersuchung des ISD hat in den Wochen vor den Zwischenwahlen in den USA im November 2022 wahlbezogene Desinformation auf YouTube, Instagram und TikTok analysiert und dabei nachgewiesen, dass diese Online-Plattformen es versäumt haben, geeignete Vorkehrungen zur Berücksichtigung des Einsatzes von Kurzvideos als Instrument zur Verbreitung von Desinformation bei Wahlen zu treffen. Siehe: Martiny, M., Jones, I. & Cooper, L. (2022). Election disinformation thrives following social media platforms' shift to short-form video content. ISD. Abrufbar unter: https://www.isdglobal.org/digital_dispatches/election-disinformation-thrives-following-social-media-platforms-shift-to-short-form-video-content/

8 Chung, A. (2019). News Feeds, Old Content: A Brief History of Algorithmically Curated Feeds on Facebook and Twitter. *Medium*. Abrufbar unter: https://medium.com/@annawchung/news-feeds-old-content-a-brief-history-of-algorithmically-curated-feeds-on-facebook-and-twitter-85b5e5d8e30a

9 Mark Zuckerberg (12. Januar 2018). Status update. *Facebook*. Abrufbar unter: https://www.facebook.com/zuck/posts/10104413015393571#

10 Mosseri, A. (2018). Bringing People Closer Together. *Meta*. Abrufbar unter: https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/

11 Nicht für die Öffentlichkeit bestimmtes Dokument, das CNN vorliegt. Abrufbar unter: https://s3.documentcloud.org/documents/21093256/internal-document-obtained-by-cnn.pdf

12 Lada, A., Wang, M., & Yan, T. (2021). How does news feed predict what you want to see? Facebook Newsroom. *Meta*. Abrufbar unter: https://about.fb.com/news/2021/01/how-does-news-feed-predict-what-you-want-to-see/

13 YouTube (2022). Analytics and Reporting APIs. Metrics. *YouTube*. Abrufbar unter: https://developers.google.com/youtube/analytics/metrics

14 Meta (2022). Content borderline to the Community Standard. *Meta Transparency Center*. Abrufbar unter: https://transparency.fb.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards

15 Patel, N. (13. September 2022). Everyone knows what YouTube is – few know how it really works. *The Verge*. Abrufbar unter: https://www.theverge.com/2022/9/13/23349037/mark-bergen-youtube-creators-tiktok-algorithm

16 Metz, R. (27. Oktober 2021). Likes, anger emojis and RSVPs: the math behind Facebook's News Feed — and how it backfired. *CNN Business*. Abrufbar unter: https://edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math/index.html

17 Hindman, M., Lubin, N. and Davis, T. (10. Februar 2022). Facebook Has a Superuser-Supremacy Problem. *The Atlantic*. Abrufbar unter: https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/

18 Edelman, G. (19. November 2021). How Facebook Could Break Free From the Engagement Trap. *Wired*. Abrufbar unter: https://www.wired.com/story/jeff-allen-interview-facebook-engagement-trap/

19 Davidson, T., Bhattacharya, D. und Weber, I. (2019) Racial Bias in Hate Speech and Abusive Language Detection Datasets. *arXiv*. Abrufbar unter: https://arxiv.org/pdf/1905.12516.pdf

20 Noble, S. (2018). Algorithms of oppression. Databite No. 109. *Data & Society*. Abrufbar unter: https://datasociety.net/library/safiya-umoja-noble-algorithms-of-oppression/

21 Simonite, T. (25. Oktober 2021) Facebook Is Everywhere; Its Moderation Is Nowhere Close. *Wired*. Abrufbar unter: https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

22 Ovadya, A. (2022). Bridging-Based Ranking. *Belfer Center*. Abrufbar unter: https://www.belfercenter.org/publication/bridging-based-ranking

23 Chung, A. (2019). News Feeds, Old Content: A Brief History of Algorithmically Curated Feeds on Facebook and Twitter. *Medium*. Abrufbar unter: https://medium.com/@annawchung/news-feeds-old-content-a-brief-history-of-algorithmically-curated-feeds-on-facebook-and-twitter-85b5e5d8e30a

24 O'Neil, C. (2017) *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

25 Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of Machine Learning Research 81:1–15, *2018 Conference on Fairness, Accountability, and Transparency*. Abrufbar unter: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

26 Ada Lovelace Institute (2021) Technical methods for regulatory inspection of algorithmic systems. *Ada Lovelace Institute*. Abrufbar unter: https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

27 Official Journal of the EU (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (Digital Services Act). Volume 65. 27 Oktober 2022. Abrufbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2022:277:FULL&from=EN

28 Nähere Informationen über das Europäische Zentrum für Algorithmentransparenz (European Centre for Algorithmic Transparency, ECAT) finden Sie auf der Website der Europäischen Kommission unter: https://digital-strategy.ec.europa.eu/en/policies/ecat [Zugriff am 31. Oktober 2022]

29 Vermeulen, M. (2022). Researcher Access to Platform Data: European Developments. *Journal of Online Trust and Safety*. Vol. 1 No. 4 (2022). Abrufbar unter: https://tsjournal.org/index.php/jots/article/view/84

30 European Commission (2022). 2022 Strengthened Code of Practice on Disinformation. *European Commission*. Abrufbar unter: https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

31 In dem gestärkten Verhaltenskodex zur Bekämpfung von Desinformation schließt »Desinformation« die Konzepte »Fehlinformation«, »Desinformation«, »Einflussnahme auf Informationen« sowie »Einmischungen aus dem Ausland in den Informationsraum« ein, die im Europäischen Aktionsplan für Demokratie der Kommission definiert werden (S. 18). Abrufbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

32 Vermeulen, M. (2022). Researcher Access to Platform Data: European Developments. Journal of Online Trust and Safety. Vol. 1 No. 4 (2022). Abrufbar unter: https://tsjournal.org/index.php/jots/article/view/84

33 Guhl, J., Marsh, O. und Tuck, H. (2022). Researching the Evolving Online Ecosystem: Barriers, Methods and Future Challenges. *ISD*. Abrufbar unter: https://www.isdglobal.org/wp-content/uploads/2022/07/Researching-the-Evolving-Online-Ecosystem_Main-report.pdf

34 Roumeliotis, G. (4. November 2019). Exclusive: U.S. Opens National Security Investigation into TikTok – Sources. *Reuters*. Abrufbar unter: https://www.reuters.com/article/us-tiktok-cfius-exclusive-idUSKBN1XB4IL

35 O'Connor, C. (2021) Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok. *ISD*. Abrufbar unter: https://www.isdglobal.org/wp-content/uploads/2021/08/HateScape_v5.pdf

36 Camargo, C. Q. und Simon, F. M. (2022). Mis- and disinformation studies are too big to fail: Six suggestions for the field's future. *Harvard Kennedy School Misinformation Review*. September 2022, Volume 3, Issue 5. Abrufbar unter: https://misinforeview.hks.harvard.edu/article/mis-and-disinformation-studies-are-too-big-to-fail-six-suggestions-for-the-fields-future/

37 Vgl. demnächst erscheinendes Paper: Meßmer, A.-K. and Degeling, M. (202x). Auditing Recommender Systems with Risk Cards. *Stiftung Neue Verantwortung*. Abrufbar unter: https://www.stiftung-nv.de/en/subproject/approaches-analyse-and-evaluate-ai-based-recommendation-systems-internet-intermediaries

38 Dimson, T. (2022). How Recommendation Algorithms Actually Work. *Future*. Abrufbar unter: https://future.com/forget-open-source-algorithms-focus-on-experiments-instead/

39 Christchurch Call (2022). *Christchurch Call Initiative on Algorithmic Outcomes*. Abrufbar unter: https://www.christchurchcall.com/media-and-resources/news-and-updates/christchurch-call-initiative-on-algorithmic-outcomes/

40 White House (2022). *Blueprint for an AI Bill of Rights*. Abrufbar unter: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

41 Masnick, M. (2019) Protocols, Not Platforms: A Technological Approach to Free Speech. *Knight First Amendment Institute*. Abrufbar unter: https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

42 Gobo.social. *MIT Media Lab's Center for Civic Media*. Abrufbar unter: https://www.media.mit.edu/projects/gobo/overview/

43 Block Party. Abrufbar unter: https://www.blockpartyapp.com/

44 Twitter (2022). Twitter Toolbox. *Developer Platform*. Abrufbar unter: https://developer.twitter.com/en/community/toolbox

45 Perspective (2022). Abrufbar unter: https://developers.perspectiveapi.com/s/about-the-api

46 Keller, D. (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy*. Vol. 32. Issue 3. 168-72. Abrufbar unter: https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/

47 Marsh, O. (2022). Social Media Futures: Interventions Against Online Unpleasantness. *Tony Blair Institute for Global Change*. Abrufbar unter: https://institute.global/policy/social-media-futures-interventions-against-online-unpleasantness

48 Keller, D. (2021). The Future of Platform Power: Making Middleware Work. *Journal of Democracy*. Vol. 32. Issue 3. 168-72. Abrufbar unter: https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/

49 OECD (2021). Data portability, interoperability and digital platform competition. *OECD Competition Committee Discussion Paper*. Abrufbar unter: http://oe.cd/dpic

50 Fukuyama, F. et al.. Middleware for Dominant Digital Platforms: A Technological Solution to a Threat to Democracy. *Stanford Cyber Policy Center*. 3. Abrufbar unter: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf

51 BIT, CDEI and Doteveryone (2022). Active Online Choices: Designing to Empower Users. *Centre for Data Ethics and Innovation (CDEI)*. Abrufbar unter: https://www.bi.team/wp-content/uploads/2021/08/CDEI-Active-Online-Choices_Final-Report-1.pdf

52 eSafety Commissioner (2021). Development of industry codes under the Online Safety Act. Position Paper. *eSafety Commissioner*. Abrufbar unter: https://www.esafety.gov.au/sites/default/files/2021-09/eSafety%20Industry%20Codes%20Position%20Paper.pdf

53 eSafety Commissioner (2021). Online Safety Act 2021. Fact sheet. *eSafety Commissioner*. Abrufbar unter: https://www.esafety.gov.au/sites/default/files/2021-07/Online%20Safety%20Act%20-%20Fact%20sheet.pdf

54 Bush, D. und Zaheer, A. (2019) Bing's Top Search Results Contain an Alarming Amount of Disinformation. Freeman Spogli Institute for International Studies. Stanford University. Abrufbar unter: https://fsi.stanford.edu/news/bing-search-disinformation

55 Meta (2019). People, Publishers, the Community. *Facebook Newsroom*. Abrufbar unter: https://about.fb.com/news/2019/04/people-publishers-the-community/

56 Meta (2022). Arten von Inhalten, die wir herabstufen. *Transparency Center*. Abrufbar unter: https://transparency.fb.com/de-de/features/approach-to-ranking/types-of-content-we-demote/

57 Smith, J., Leavitt, A. und Jackson, G. (2018). Designing New Ways to Give Context to News Stories. *Facebook Newsroom*. Abrufbar unter: https://about.fb.com/news/2018/04/inside-feed-article-context/

58 Singh, S. (2021). Facebook Releases Information on Algorithmic Content Ranking, but More Transparency Is Needed. Open Technology Institute. *New America*. Abrufbar unter: https://www.newamerica.org/oti/blog/facebook-releases-information-on-algorithmic-content-ranking-but-more-transparency-is-needed/

59 Google (2019). How Google Fights Disinformation. *Google*. S.12. Abrufbar unter: https://kstatic.googleusercontent.com/files/388aa7d18189665e5f5579aef18e181c2d4283fb7b0d4691689dfd1bf92f7ac2ea6816e09c02eb98d5501b8e5705ead65af653cdf94071c47361821e362da55b

60 Ibid, S.4, S.12.

61 Google (2022). Quality Raters. General Guidelines, S. 23. *Google*. Abrufbar unter: https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

62 Chang, J. (2017). Identifying credible content online, with help from the Trust Project. *Google Keyword*. Abrufbar unter: https://blog.google/outreach-initiatives/google-news-initiative/sorting-through-information-help-trust-project/

63 Google (2019) What site owners should know about Google's August 2019 core update. *Google Search Central Blog*. Abrufbar unter: https://developers.google.com/search/blog/2019/08/core-updates

64 Google Patents. Producing a ranking for pages using distances in a web-link graph. *Google Patents*. Abrufbar unter: https://patents.google.com/patent/US9165040B1/en

65 Allen, J. (2021). The Integrity Institute's Analysis of Facebook's Widely Viewed Content Report (Q4 2021 - Q1 2022). *Integrity Institute*. Abrufbar unter: https://integrityinstitute.org/widely-viewed-content-analysis-tracking-dashboard

66 Singh, S. (2019). Social media is broken. A new report offers 25 ways to fix it. *MIT Sloan School of Management*. Abrufbar unter: https://mitsloan.mit.edu/ideas-made-to-matter/social-media-broken-a-new-report-offers-25-ways-to-fix-it

67 Parliament of Australia (2022). Select Committee on Social Media and Online Safety - 03/02/2022 – Online harms that may be faced by Australians on social media and other online platforms. *Parliamentary Business*. Abrufbar unter: https://www.aph.gov.au/Parliamentary_Business/Hansard/Hansard_Display?bid=committees/commrep/25635/&sid=0000

68 Walsh, D. (2020). Study: 'Accuracy nudge' could curtail COVID-19 misinformation online. *MIT Sloan School of Management*. Abrufbar unter: https://mitsloan.mit.edu/ideas-made-to-matter/study-accuracy-nudge-could-curtail-covid-19-misinformation-online

69 Congress.gov (2022). S.3608 - Social Media NUDGE Act. *Congress.gov*. Abrufbar unter: https://www.congress.gov/bill/117th-congress/senate-bill/3608?s=1&r=5

70 Goodman, E.P. (2022). Assessing the NUDGE Act. *Tech Policy Press*. Abrufbar unter: https://techpolicy.press/assessing-the-nudge-act/

71 Birdwatch (2022). Note ranking. *Twitter, Inc.*. Abrufbar unter: https://twitter.github.io/communitynotes/ranking-notes/

72 Birdwatch (2022). Matrix Factorization. *Twitter, Inc.*. Abrufbar unter: https://twitter.github.io/communitynotes/ranking-notes/#matrix-factorization

73 Coleman, K. (2021). Introducing Birdwatch, a community-based approach to misinformation. *Twitter Inc.*. Abrufbar unter: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

74 Birdwatch (2022). Helpful Birdwatch notes are now visible to everyone on Twitter in the US. *Twitter Inc.*. Abrufbar unter: https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us

75 Ovadya, A. (2022). Bridging-Based Ranking. *Belfer Center*. Abrufbar unter: https://www.belfercenter.org/publication/bridging-based-ranking

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdgermany.org

gefördert durch:
Auswärtiges Amt