HEINRICH BÖLL STIFTUNG
GUNDA WERNER INSTITUT

E-PAPER

Misogynoir – die algorithmische Diskriminierung Schwarzer Frauen in der Content-Moderation

VON BRANDEIS MARSHALL

Herausgegeben von der Heinrich-Böll-Stiftung, Oktober 2022

Misogynoir – die algorithmische Diskriminierung Schwarzer Frauen in der Content-Moderation

Von Brandeis Marshall

Inhaltsverzeichnis

Vorwort 3

1 Einleitung 4

2 Die Content-Moderation in der aktuellen Praxis 6

2.1 Das Problem der Verallgemeinerung 6

2.2 Mit zweierlei Maß gemessen 7

3 Vorschläge und Anregungen 12

3.1 Die Rolle der Social-Media-Unternehmen 12

3.2 Umgang mit strukturellen Ungleichheiten 13

3.3 Asymmetrische Machtverhältnisse zwischen verfassenden und kommentierenden Personen ausgleichen 15

Literaturverzeichnis 17

Die Autorin 21

Vorwort

Die sozialen Medien sind in unserer Gesellschaft als Ort des Meinungsaustausches und der Wissensvermittlung tief verankert und haben sich zu einem politischen Diskurs-, aber auch Dissonanz-Raum entwickelt. Deshalb ist die Frage, wer Zugang zu diesen Räumen hat, essentiell im Sinne demokratischer Beteiligung. Aus einer intersektional feministischen Perspektive, die die Heinrich-Böll-Stiftung etwa im Gunda-Werner-Institut institutionalisiert hat, ergeben sich Fragen nach der Struktur von Exklusion und der systematischen Unsichtbarmachung von User*innen.

In der vorliegenden Publikation untersucht die US-Autorin den Einfluss rassischer, misogyner Diskriminierungen, welche insbesondere zum Ausschluss Schwarzer Frauen führen. Dies ist eine notwendige Intervention, wenn es um die Frage der Gestaltung der Sozialen Medien für eine geschlechtergerechte demokratische Zukunft geht.

Berlin, im Oktober 2022

Katharina Klappheck
Gunda-Werner-Institut für Feminismus & Geschlechterdemokratie
*Referent*in für feministische Netzpolitik*

1 Einleitung

Wir alle gehören der Gattung Mensch an. Dennoch zwingen soziale, wirtschaftliche und politische Konstrukte uns eine Hierarchie auf, in deren Mittelpunkt der weiße Kolonialismus und das weiße Patriarchat stehen – verbunden mit Manipulation, Zwang oder auch Gewalt. Diese sozialen, wirtschaftlichen und politischen Strukturen orientieren sich nicht am Maßstab der Gerechtigkeit. Vielmehr sind sie darauf ausgerichtet, alles außerhalb der weißen Kultur abzuwerten. Wie Gulati-Partee und Potapchuk in ihrer Arbeit zur *Förderung der Gleichstellung aller ethnischen Gruppen* in den USA darlegen, werden «ethnische Unterschiede durch politische Maßnahmen des öffentlichen und privaten Sektors gefördert und aufrechterhalten, wodurch nicht nur die Communities of Color benachteiligt, sondern Weiße zudem übervorteilt werden».

Unsere globalisierte Gesellschaft strahlt Ablehnung gegenüber Schwarzen Menschen aus – und das gilt auch für das Internet. Die Technologiebranche wird in den Bereichen *Unternehmertum, Führung* und *Belegschaft* von weißen Männern monopolisiert. Schwarze Frauen erhalten bei Unternehmensgründungen 20-mal weniger Fördermittel, verglichen mit dem üblichen Mittelwert in den USA, der Unternehmerinnen und Unternehmer aller übrigen Bevölkerungsgruppen einschließt. Dies zeigt, wie viel schwieriger es für Schwarze Frauen ist, das benötigte Kapital zu beschaffen. Im Technologiesektor sind 70-80 Prozent der Führungs- und Arbeitskräfte männlich. Diese Branche hat eine umfassend belegte toxische Unternehmenskultur aufgebaut, in der *Männer alle Karten in der Hand halten sowie sämtliche Regeln aufstellen* und *in der ein verdeckter struktureller Rassismus fortbesteht*. Diese toxische Tech-Kultur wird sich zudem so lange fortsetzen, wie sich *Lehrkräfte im Fach Informatik nicht für eine Veränderung einsetzen*. Diese Situation macht das Ringen um mehr Gleichstellung nicht gerade einfacher.

Eine Schwarze Frau zu sein bedeutet, zwei Extreme zu erleben – sowohl offline als auch online: Erstens, niemand achtet auf das, was du sagst oder wie du es sagst. Und zweitens, deine Worte sind ein gefundenes Fressen und werden penibel auseinandergenommen, überwacht und verurteilt. Unsichtbar zu sein und gleichzeitig auf dem Präsentierteller zu stehen – das ist ein stets präsentes Thema (wie in *Thick* von Tressie McMillian Cottom untersucht) sowie teilweise das Produkt der Haltung, die Moya Bailey «Misogynoir» nennt (*Crunk Feminist Collective, März 2010*) oder die «gegen Schwarze Menschen gerichtete rassistische Misogynie, die Schwarze Frauen erfahren». *The Abuse and Misogynoir Playbook* spricht von einen Fünf-Phasen-Zyklus, der Ungläubigkeit, Abwertung und Diskreditierung der Wortbeiträge Schwarzer Frauen als historischen Normalfall konstatiert. Die Definition der algorithmischen Misogynoir leitet Bailey aus ihrer Darstellung der Folgen ab, die diese Erfahrungen online und verschlüsselt für Schwarze Frauen haben.

In der digitalen Welt sind wir Schwarze Frauen mit unserer Präsenz, unseren Erfahrungen und unseren Interaktionen breit vertreten – von der Verkündung unserer Erfolge bis hin

zum Austausch über unsere Traumata. Doch inzwischen kommen Reaktionen und Antworten extrem schnell, von echten Profilen, aber auch von Bot- und Troll-Profilen aus der ganzen Welt.

Schwarze Menschen im Allgemeinen und Schwarze Frauen im Besonderen sind *häufiger Zielscheiben von Belästigung im Netz* als weiße Menschen. Es hat Fälle Schwarzer Frauen gegeben, die international aufgrund ihrer Gelehrsamkeit kritisiert wurden. Ihre Posts wurden entfernt, weil ihre Äußerungen viel genauer unter die Lupe genommen wurden als die ihrer weißen Kolleginnen. Ihre Accounts in den Sozialen Netzwerken wurden temporär deaktiviert oder sogar gelöscht, weil sie sich gegen jegliche Form der algorithmischen Diskriminierung ausgesprochen hatten.

Die besondere Situation der Schwarzen Frauen wird weder anerkannt noch thematisiert. Der Bericht *Stop Online Violence Against Black Women von Shireen Mitchell aus dem Jahr 2018* zeigt, wie Online-Kampagnen mithilfe von Facebook-Anzeigen erstellt wurden, um Schwarze Mädchen und Frauen vor und während der US-Präsidentschaftswahl 2016 mit sexualisierten Memes, Hashtags und gefälschten Konten zu diskreditieren und Fehlinformationen zu verbreiten. Charisse C. Levchek dokumentiert darüber hinaus in *Microaggressions and Modern Racism* den gegen Schwarze Menschen gerichteten Rassismus auf der Mikro- und Makroebene, der sowohl über persönliche Interaktionen als auch virtuell stattfindet. Levcheck verweist auf rassistische Verunglimpfungen und andere Formen von Hate Speech, die auf der Mikro- und Makro-Aggressionsebene im Internet zirkulieren. Sie ruft insbesondere Unternehmen dazu auf, Strategien und Verfahren zur Bekämpfung von Rassismus einzuführen, Mikro- und Makro-Aggressionen zu bestrafen und Opfer rassistischer Gewalt zu unterstützen.

Matamoros-Fernandez spricht in diesem Zusammenhang von «plattformbasiertem Rassismus», einer «neuen Form von Rassismus, die sich aus der Kultur der Sozialen Netzwerke ableitet – aus ihrer Struktur, ihren technischen Möglichkeiten, Geschäftsmodellen und Regeln – sowie den spezifischen Nutzungskulturen, die mit ihnen verbunden sind». Die Moderation von Inhalten in den sozialen Netzen könnte durchaus zur Schaffung integrativer, einladender Räume für Schwarze Frauen beitragen. Doch die derzeitige Praxis unterstützt die Misogynoir und setzt sie auch in den Algorithmen um.

2 Die Content-Moderation in der aktuellen Praxis

2.1 Das Problem der Verallgemeinerung

Nach *Grimmelmann* unterliegt die Content-Moderation den «Governance-Mechanismen, die die Teilnahme an einer Gemeinschaft strukturieren, um die Zusammenarbeit zu erleichtern und Missbrauch zu verhindern».

Abb. 1: So funktioniert die Content-Moderation

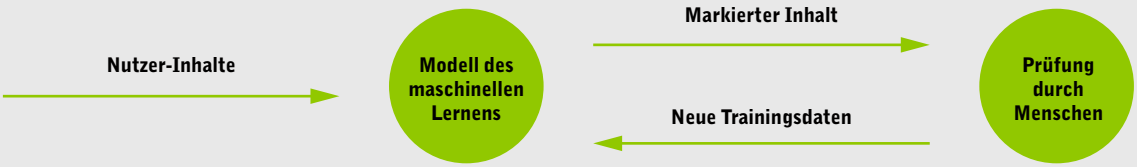


Abbildung 1 zeigt die Grundpfeiler der Content-Moderation. Von Nutzenden erstellte Inhalte gelangen in ein Soziales Netzwerk oder ein anderes digitales System, in dem eine Reihe von Algorithmen für maschinelles Lernen ausgeführt wird, um die Angemessenheit der Inhalte automatisch zu prüfen und zu bewerten.

Solange ein Beitrag den Governance-Regeln des Netzwerks entspricht, wird er auch veröffentlicht. Verstößt der Post jedoch dagegen, wird er als sensibler Inhalt gekennzeichnet, nicht veröffentlicht oder gleich entfernt. Als potenziell problematisch eingestufte Inhalte werden an menschliche Content-Moderatorinnen und -moderatoren weitergeleitet, die eine manuelle Prüfung vornehmen und entscheiden, ob sie veröffentlicht werden können. Das Problem bei der Content-Moderation in den Sozialen Netzwerken ist die Bewältigung des schnellen Zustroms an nutzergenerierten Inhalten – ein Teufelskreis, in dem mehr Daten zu immer mehr Daten führen.

Wie in *Gillespies* Buch *Custodians of the Internet* beschrieben, hat die Kontrolle von Daten und Inhalten inzwischen höchste Priorität für die Sozialen Netzwerke. Sie setzen hierfür eine Kombination aus menschlicher Bearbeitung und automatisierten Verfahren zur Content-Moderation ein. Angesichts des überschaubaren Umfangs an nutzergenerierten Inhalten in den Anfängen des Internets wurde die Überprüfung durch Menschen zunächst als ausreichend angesehen. Das manuelle Screening konnte jedoch bald nicht mehr mit dem Tempo der nutzergenerierten Inhalte Schritt halten. Es wurden also automatisierte Verfahren zur Content-Moderation entwickelt. Hierdurch sollte die zusätzliche Arbeitsbelastung bewältigt und die Moderation, zumindest theoretisch, effizienter werden.

Besonders anstößige Inhalte, wie z. B. Kinderpornografie, bedeuten eine extrem hohe psychische Belastung für die menschlichen Content-Moderatorinnen und -moderatoren. Automatisierte Algorithmen zur Content-Moderation können Menschen davor schützen, diese Inhalte manuell überprüfen zu müssen.

(Nebenbei bemerkt: Facebook und Twitter standen bereits mehrfach wegen arbeitsrechtlicher Probleme im Zusammenhang mit ihrer Content-Moderation in der Kritik. Die Moderatorinnen und Moderatoren von Facebook beklagten *einen Mangel an psychologischer Unterstützung* und *insgesamt nicht zufriedenstellende Arbeitsbedingungen*. In *Paul M. Barretts Bericht zur Moderation von Inhalten und deren Folgen von 2020* wird Facebook unter anderem dazu aufgerufen, Content-Moderatorinnen und -moderatoren fest einzustellen und die Zahl der Mitarbeitenden zu verdreifachen, anstatt die anfallende Mehrarbeit an externe Dritte zu vergeben).

Dennoch bleibt die Content-Moderation ein schwieriges Thema. Viele Inhalte können von den Algorithmen nicht richtig kategorisiert werden, da die Bedeutungsnuancen unserer Kultur doch nicht so vorhersehbar sind wie einst angenommen. Schwarze Frauen sind sowohl von sozialen, wirtschaftlichen und politischen Faktoren als auch von technischen/algorithmischen, geschlechtsspezifischen und ethnischen Aspekten betroffen. Der übliche gesellschaftliche Ansatz betrachtet diese Faktoren in der Regel getrennt voneinander bzw. geht einen nach dem anderen an in der Annahme, die Lösung einzelner Teilprobleme werde zur Lösung der Gesamtproblematik führen.

In der Tendenz werden die schwierigsten Themen über einen Kamm geschoren. Ein- und dieselbe Maßnahme soll die unterschiedlichen Problemstellungen gleichermaßen abdecken. Die standardisierte Erkennung und Verarbeitung der angezeigten Inhalte, insbesondere der problematischen, ist nur über die Festlegung von Gleichheitskriterien möglich. Diese ermitteln unangemessene Inhalte über konsistente Muster. Außerdem funktioniert dieses Verfahren nur über eine universelle Routine zur Problemlösung – was in unserer globalen Gesellschaft, die in ihrer demografischen Zusammensetzung eher multikulturell ist, nicht funktionieren kann.

Denn das Problem ist nicht die Addition disparater Effekte, sondern eher deren Multiplikation. Das Ergebnis dieser Vereinfachung ist eine inkonsistente Praxis der Content-Moderation, die Schwarze Frauen letztendlich zum Schweigen bringt.

2.2 Mit zweierlei Maß gemessen

Nominell unterliegt die Content-Moderation den Standards und Regeln der Sozialen Netzwerke, etwa *den Community Standards von Facebook*. Facebook definiert in seiner Regel zur Content-Moderation 23 verschiedene Kategorien, unter anderem Gewalt und kriminelles Verhalten, Sicherheit, anstößige Inhalte, Integrität und Authentizität,

geistiges Eigentum sowie inhaltsbezogene Anfragen und Entscheidungen. Facebook gibt ausführlich an, welche Inhalte als unangemessen gelten, allerdings nicht, wie die Standards umgesetzt werden.

So wurde beispielsweise ein Facebook-Kommentar der Highschool-Lehrerin *Carolyn Wysinger* aus Richmond (Kalifornien) innerhalb von 15 Minuten gelöscht. Die Begründung war Hate Speech. Die Lehrerin hatte auf einen Post des Hollywood-Schauspielers Liam Neeson reagiert, der sich gegen Schwarze Menschen richtete. Während er einen Rachefilm promotete, hatte Neeson zugegeben, dass er Jahrzehnte zuvor, nach der Vergewaltigung einer Freundin durch einen ihr unbekannten Schwarzen, «*durch die Straßen gezogen sei, um Schwarze zu jagen und anzugreifen*».

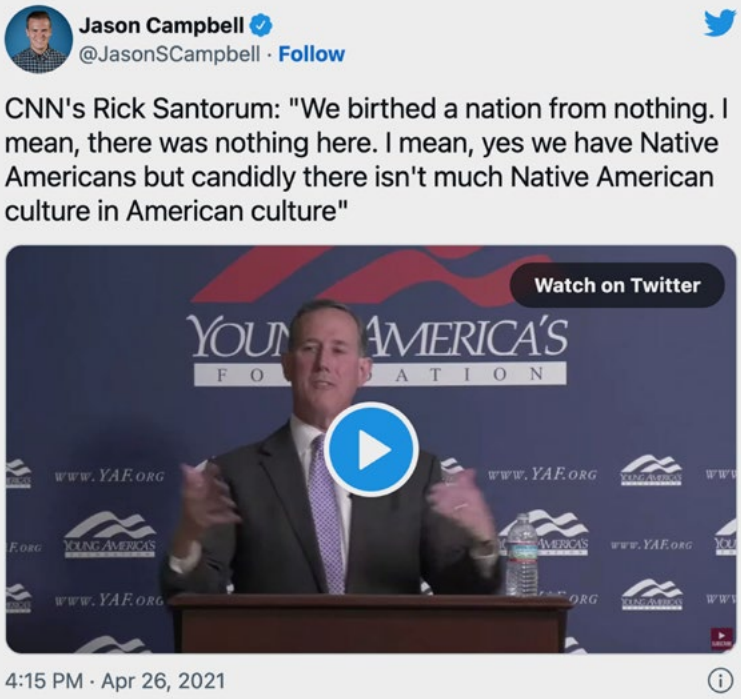
Der Post des Schauspielers wurde nicht entfernt. Die Darstellung von Gewalt gegen Schwarze Männer auf der Grundlage ihrer Hautfarbe zählt offenbar also nicht zu den löschbedürftigen Inhalten. Einen Weißen hingegen als «fragil» zu bezeichnen, ist offenbar bereits Hate Speech. Den 23 Kategorien von Facebook zufolge verstößt Liam Neesons Post sowohl gegen die Regeln für Gewalt und Aufhetzung als auch für Hate Speech. Es ist unklar, gegen welche Regel Wysinger mit ihrem Beitrag verstoßen hat.

Frauen werden regelmäßig als fragil, schwach und empfindlich bezeichnet. Solche Posts werden jedoch nicht als Hate Speech eingeordnet. War es also vielleicht das Wort «fragil» in Verbindung mit «weißen Männern», das die automatischen Algorithmen zur Content-Moderation auslöste? Dies ist nicht mit Bestimmtheit festzustellen. Denn in puncto Algorithmen, Prozesse und Praktiken in diesem Bereich gibt es keine Transparenz. Einblicke bietet hier in erster Linie das Dokument *Facebook's Community Standards Enforcement Report*. Der stark bereinigte Bericht präsentiert die Daten allerdings in aggregierter und zusammengefasster Form.

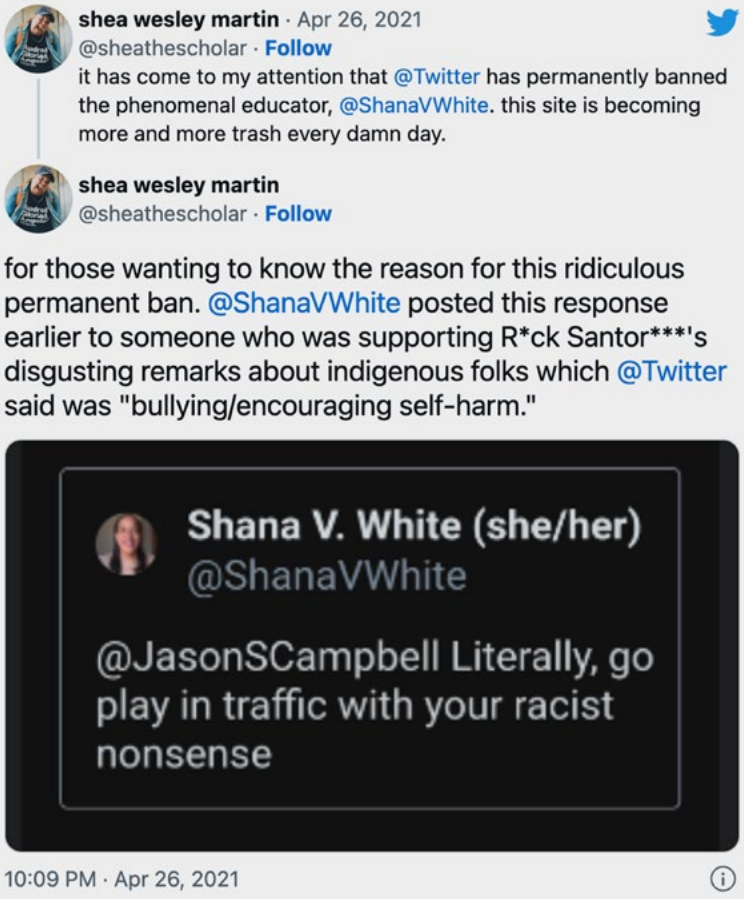
Bei Twitter fallen *15 Regelkategorien* unter den Begriff der Content-Moderation, unterteilt in vier Gruppen: Sicherheit, Datenschutz, Authentizität und Durchsetzung sowie Einsprüche. In der Content-Moderation fallen Entscheidungen in Bezug auf die Posts Schwarzer Frauen unverzüglich und mit aller Härte.

Betrachten wir beispielhaft den Fall von *Shana V White*, derzeit Senior Associate der CS Equity and Justice Initiatives am Kapor Center. Als Informatiklehrerin mit 16 Jahren Berufserfahrung engagiert sie sich bei ihren über 25.000 Twitter-Followern aktiv für Lehrerinnen und Lehrer sowie für ausgegrenzte Gruppen. White wurde am 26. April 2021 *dauerhaft von der Plattform verbannt*, nachdem sie auf den Post einer Person *reagiert hatte*, der die Falschinformationen des ehemaligen Senators Rick Santorum zur indigenen Bevölkerung Nordamerikas unterstützte. Sie legte Einspruch ein, und ihr Konto wurde noch am selben Abend wieder aktiviert. Einige Tage später postete White erneut Beiträge zur Unterstützung einer marginalisierten Gruppe; ihr Konto wurde dauerhaft gesperrt. Am 27. Mai 2021 *kündigte sie ihren Umstieg auf die Plattform Twitter an*.

Jason S Campbells Tweet vom 26. April 2021



Antwort-Tweet von Shana V. White auf den Post vom 26. April 2021



Ein Post, der Falschinformationen zur Geschichte der Vereinigten Staaten verbreitet, bleibt also dauerhaft stehen (~9,8 Millionen Aufrufe), aber ein lediglich bissiger, sarkastischer Kommentar führt gleich zu einer dauerhaften Sperrung des Profils? Der ursprüngliche Post verstößt augenscheinlich gegen die Regel zu Manipulation der Plattform und Spam.

Whites Reaktion verstieß wahrscheinlich gegen die Regel zu Selbstmord und Selbstverletzung. Doch auch hier ist eine bestimmte Einordnung nicht möglich, da es an der erforderlichen Transparenz mangelt. Wie Facebook ist auch der *Transparenzbericht von Twitter* stark bereinigt und präsentiert lediglich aggregierte und zusammengefasste Daten.

Schwarze Frauen werden in diesem Dickicht der Content-Moderation zur Zielscheibe. Sie reagieren auf hetzerische Posts und werden doch immer wieder zum Schweigen gebracht. Wenn Schwarze Frauen sich selbst oder andere verteidigen, werden sie als Aufwieglerinnen gemeldet, die von den einflussreichen Kräften des weißen Kolonialismus in die Schranken verwiesen werden müssen.

Wir wissen nicht, inwieweit diese algorithmische Diskriminierung Schwarzer Frauen durch die Content-Moderationsalgorithmen selbst oder durch andere Nutzende ausgelöst wird, die die Beiträge Schwarzer Frauen melden. In jedem Fall fließt die endgültige Entscheidung über die Angemessenheit ihrer Posts – unabhängig davon, wie sie ausfällt – in das Verzeichnis der verdächtigten und als bestätigt gekennzeichneten Regelverstöße der Plattform ein (*siehe den Abschnitt zu Konsequenzen bei Verstößen gegen die Regel zum Umgang mit Hasskommentaren bei Twitter*). Der Text mutet wie eine Kopie des US-amerikanischen Strafrechts an, übertragen in die digitale Welt: Hohe Hürden stehen der Aufhebung eines einmal ausgesprochenen Verdachts der Hetze gegen eine Person entgegen. Twitter sieht *mehrere Schritte der Content-Moderation* vor – sowohl auf Tweet-, Direktnachrichten- als auch auf Account-Ebene, wobei die dauerhafte Sperrung die schwerwiegendste Konsequenz darstellt. Der Eskalationsprozess für die Durchsetzung der Content-Moderation erscheint dennoch weiterhin unklar.

Nachfolgend führen wir weitere öffentlichkeitswirksame Fälle von Content-Moderation und digitaler Gewalt auf:

- Dr. Safiya Nobles Buch *Algorithms of Oppression: How Search Engines Reinforce Racism* wurde von einem Historiker des internationalen Berufsverbands IEEE (Institute of Electrical and Electronics Engineers) *öffentlich auf Twitter* kritisiert. Dieser hatte das Buch nicht einmal gelesen, bevor er seinen Kommentar im Namen des IEEE postete. Der Beitrag wurde erst nach öffentlichen Protesten entfernt.
- Die *Gender Shades-Studie* von Joy Buolamwini, Dr. Timnit Gebru, Dr. Helen Raynham und Deborah Raji, die rassistische und geschlechtsspezifische Diskriminierung in kommerzieller Gesichtserkennungssoftware untersucht, wurde von *Amazon* heftig attackiert – das Unternehmen hat eine der in der Studie behandelten Technologien

entwickelt. Amazon ging sogar so weit, die Kritik an den Ergebnissen der Studie *in seinem Blog* zu veröffentlichen. Buolamwini reagierte, indem sie ihre älteren Posts an die einzelnen Unternehmen, in denen sie den Rassismus und die Genderdiskriminierung der Gesichtserkennung auf der Publishing-Plattform *Medium* teilte.

– Dr. Timit Gebru *erfuhr digitale Gewalt*, als sie sich zum *an der Duke University entwickelten KI-Tool PULSE* zur Erstellung menschlicher Gesichter äußerte. Ihre bahnbrechende Forschung auf dem Gebiet der algorithmischen Ungleichheiten in der künstlichen Intelligenz wurde heruntergespielt, sie persönlich wurde als übertrieben emotional, sogar als hysterisch bezeichnet, musste *Mansplaining* ertragen und wurde dem rassistischen Stereotyp der *wütenden Schwarzen Frau* entsprechend behandelt.

– Einige Monate später plädierte Dr. Gebru für mehr Transparenz bei den internen Google-Prozessen in Bezug auf die Veröffentlichungskriterien für Forschungsarbeiten. Eines ihrer Papers, das bei namhaften internationalen Konferenz zur Präsentation ausgewählt worden war, wurde von Google genauestens geprüft. Sie erhielt jedoch keinerlei Rückmeldung zum Stand der Bearbeitung und zu den Gründen der Beanstandung. Dr. Gebrus Account wurde sofort gesperrt. Als die Wissenschaftlerin das Ereignis in Echtzeit auf Twitter teilte, wurde sie nach der Sperrung ihres Google-Accounts *immer wieder* auch auf Twitter angegriffen. Die Accounts, von denen die Angriffe erfolgten, blieben noch monatelang aktiviert.

3 Vorschläge und Anregungen

3.1 Die Rolle der Social-Media-Unternehmen

In den USA wird die Situation der Schwarzen Frauen im Internet im Gesetzgebungsverfahren zur Content-Moderation weiterhin übersehen. Mehr als deutlich wird dies im Bericht des wissenschaftlichen Dienstes im US-Kongress zu *Sozialen Medien* (Congressional Research Service Report) vom Januar 2021, der weder die Black Communities noch Schwarze Menschen, geschweige denn Schwarze Frauen, erwähnt. In der Debatte um die Content-Moderation sticht dabei insbesondere Paragraph 230 des Communications Decency Act von 1996 hervor: «Kein Anbietender oder Nutzender eines interaktiven Computerdienstes ist als Herausgebender oder Sprechender solcher Informationen zu behandeln, die durch einem anderen Anbietenden von Inhalten bereitgestellt werden» (*47 U.S.C. § 230*). Mit anderen Worten: Interaktive Computerdienste – eine in den 1990er Jahren entstandene Bezeichnung für Internetseiten – haften nicht für Inhalte Dritter, die auf ihren Internetseiten veröffentlicht werden. Im Allgemeinen gewährleistet diese Klausel, dass Betreibende von Internetseiten nicht gerichtlich belangt werden können, wenn Nutzende illegale Inhalte einstellen. Urheberrechtsverletzungen, pornografisches Material und Verstöße gegen Bundesgesetze sind in den Ausnahmen des Abschnitts 230 ausdrücklich erwähnt.

In ihrer Funktion als Plattformen hosten, veröffentlichen und moderieren Soziale Netzwerke Inhalte, die möglicherweise zwar nicht vom Haftungsschutz ausgenommen sind, aber dennoch Schaden anrichten können, indem sie bestimmte Gruppen von Nutzenden diskriminieren. Die Regeln für die Moderation werden von wirtschaftlich und politisch einflussreichen Privatunternehmen festgelegt und unterliegen keinerlei gesetzlicher Regulierung. Der Knackpunkt der Debatte ist die Frage, wie diese Unternehmen rechtlich einzuordnen sind – ob sie als «content pass through companies» Inhalte lediglich bereitstellen oder als «responsible content companies» diese Inhalte auch verantworten – und der vage Begriff «interaktiver Computerdienst» die Social-Media-Plattformen wie Facebook und Twitter vor Haftung und Klagen aufgrund ihres Moderationsverhaltens schützt.

Sind sie als «content pass through companies» eingestuft, stellen sie nutzergenerierte Inhalte lediglich zur Verfügung (und haften auch nicht für diese). «Responsible content companies» steuern das Angebot, das den Nutzenden auf ihren Platfformen angezeigt wird (und sind entsprechend verantwortlich).

Social-Media-Plattformen gehören eindeutig zur zweiten Gruppe und sollten auch so behandelt werden. Ein Hinweis hierauf: Sie setzen automatisierte und manuelle Verfahren der Content-Moderation ein, also selbst Protokolle zur Verwaltung von Inhalten, auch

jenseits von Urheberrechtsverletzungen, pornografischem Material und Verstößen gegen US-Bundesgesetze.

Darüber hinaus muss die Definition des Begriffs «Interaktiver Computerdienst» überarbeitet werden, um beide Varianten der Interaktivität im virtuellen Raum zu erfassen. Gemäß Paragraph 230 ist ein «interaktiver Computerdienst» «jeder Anbietende von Informationsdiensten, Systemen oder Zugangssoftware, der den Computerzugriff durch mehrere Benutzende auf einen Computerserver bereitstellt, einschließlich insbesondere eines Dienstes oder Systems, das den Zugang zum Internet ermöglicht, sowie Systemen, die von Bibliotheken oder Bildungseinrichtungen betrieben oder angeboten werden» (U.S.C. §230(f)(2)).

Diese Definition umfasst diejenigen Aspekte von Computerdienst, die sich auf das Versenden von Online-Kommunikation oder Benachrichtigung beziehen, etwa die Veröffentlichung eines Artikels. Die Verarbeitung und Kontrolle der eingehenden Online-Kommunikation, z. B. von Antwortposts oder der anschließenden direkten Hin- und Her-Kommunikation zwischen Nutzenden in Bezug auf einen ursprünglichen Post, bleibt jedoch von der Definition ausgespart. Die ausgehende Online-Kommunikation ähnelt einem Rundfunksignal, welches von möglichst vielen Menschen empfangen werden soll. Die eingehende Online-Kommunikation hingegen richtet sich als Möglichkeit zum schnellen Einstieg in virtuelle Gespräche an Einzelpersonen oder Gruppen.

Die Bezeichnung «interaktiver Computerdienst» ist für die Sozialen Netzwerke also völlig unzureichend. Die Vielfalt der An- und Verwendungen von Online-Plattformen hat sich seit der Entstehung des genannten Paragraphen 230 Mitte der 1990er Jahre stark weiterentwickelt. Inzwischen findet in den Sozialen Netzwerken sowohl ausgehende als auch eingehende Kommunikation statt. Die Kommunikationsströme werden in Abhängigkeit von der Richtung jeweils unterschiedlich verarbeitet. Wie weiter unten dargestellt, verwenden Soziale Netzwerke automatisierte Algorithmen zur Content-Moderation. So können Antwortnachrichten mit höherer Priorität überwacht werden als die ausgehende Kommunikation.

3.2 Umgang mit strukturellen Ungleichheiten

Die gegenwärtige Praxis der Content-Moderation steht sinnbildlich für eine Entwicklung, die die algorithmische Misogynoir festschreibt. Der Zusammenschluss *Global Internet Forum to Counter Terrorism (GIFCT)* hat sich die «Förderung der Zusammenarbeit und des Informationsaustauschs zur Bekämpfung terroristischer und gewalttätiger extremistischer Aktivitäten im Internet» auf die Fahnen geschrieben. Die Gründungsmitglieder sind die Urgesteine der Sozialen Netzwerke und der Internet-Community: YouTube (Google), Twitter, Facebook und Microsoft.

Gorwa et al beschreiben die fachliche und politische Undurchsichtigkeit dieser Initiative sehr anschaulich. Was die Arbeitsgruppen, Partnerschaften und Kooperationen der GIFCT in Einzelnen tun, bleibt unklar. Die Transparenzberichte enthalten kaum Details darüber, welche algorithmischen Werkzeuge bei der automatisierten Content-Moderation eingesetzt werden oder welche Vorteile sich für einzelne Nutzende durch die Tätigkeit der GIFCT ergeben. Letztendlich fehlt solchen Initiativen der kontextbezogene Ansatz. Weder die Mission noch die Vision oder die Grundwerte der GIFCT enthalten Erwägungen zur ethnischen und sozialen Zugehörigkeit, Genderfragen oder Ableismus (Diskriminierung aufgrund körperlicher oder psychischer Beeinträchtigungen).

Eine Alternative bietet das *Gemeinwohlkonzept Public Value Failure (PVF)*. Es beschreibt anhand von neun Kategorien das Versagen der Gesellschaft bei der Schaffung von Gemeinwohl (Public Value). Dieses umfasst unter anderem politisch garantierte und durch entsprechende Maßnahmen gewährte Rechte, Vorteile oder Privilegien der Bürgerinnen und Bürger. Das Public-Value-Konzept ist hilfreich bei der Erarbeitung von Vorschlägen zur Herstellung von mehr Sicherheit für Schwarze Frauen im virtuellen Raum.

«Gleiche Startbedingungen» (durch breit gestreute, standardisierte Maßnahmen) sind demzufolge weniger empfehlenswert als Kollektivmaßnahmen sowie eine staatliche Politik, die sich gegen strukturelle Ungleichheit und historische Unterschiede des Zugangs zu bestimmten Möglichkeiten wendet. Es ist daher wichtig, die Content-Moderation für Schwarze Frauen und andere Minderheiten regelmäßig auf den Prüfstand zu stellen.

Die Rangfolge- und Priorisierungsprotokolle der Content-Moderation begünstigen weiße Nutzende im Vergleich zu anderen Gruppen. Schwarze Frauen werden hingegen bestraft, wenn sie auf Ungerechtigkeiten hinweisen. Die Vor- und Nachteile der Content-Moderation in ihrer gegenwärtigen Form müssen dokumentiert werden.

Die Daten sollten nach ethnischer Zugehörigkeit und Geschlecht aufgeschlüsselt werden, und zwar in sämtlichen Regeln, Standards, Praktiken, Durchsetzungsersuchen und rechtliche Anfragen zur Content-Moderation. Denn bereinigte bzw. aggregierte Daten verhindern die demografische Zuordnung mit Hinblick darauf, wer einen bestimmte Inhalt meldet, wessen Inhalte gemeldet wird, gegen welche Regel ein Beitrag verstößt und woraus die Entscheidung der automatischen Algorithmen zur Content-Moderation sich im Einzelfall ableitet. Hierbei könnte hilfreich sein, die menschlichen Moderations-Teams ausgewogen zu besetzen und Schwarzen Frauen und andere Minderheiten mit einzubeziehen. Es muss noch viel mehr getan werden, um die automatisierten Tools, Regeln und Verfahren der Content-Moderation transparent zu zeigen.

Da das Engagement in den Sozialen Netzwerken in vielen Unternehmen fester Bestandteil des Marketings ist, kann die Content-Moderation bei uneinheitlicher Anwendung dazu führen, dass Menschen um die Sicherung des Lebensunterhalts, die Wahrung ihrer Würde oder um ihr Ansehen im Internet fürchten müssen. Daher kommt auch der unabhängigen

Aufsicht eine große Bedeutung zu, neben öffentlich verfügbaren Rechenschaftsberichten und Audits. Diese Maßnahmen tragen dazu bei, Transparenz als zentralen Grundsatz der Politik in diesem Bereich zu verankern. Die gemeinsame Nutzung und Umsetzung einheitlicher Protokolle für die Content-Moderation ist Voraussetzung für echte Transparenz in allen soziopolitischen Systemen. Eine zugängliche öffentliche Debatte sowie ein reaktionsschnelles Regierungshandeln können über effektivere Kommunikationsströme Vertrauen schaffen.

3.3 Asymmetrische Machtverhältnisse zwischen verfassenden und kommentierenden Personen ausgleichen

Die Content-Moderation wirkt gleichzeitig in zwei Richtungen: (1) Interaktion der Nutzenden mit der Plattform, z. B. über die Option, andere Nutzende zu blockieren, sowie (2) Interaktion der Plattform mit dem Nutzenden – z. B. lässt die gegenwärtige Behandlung von Antwort-Tweets bestimmte Gruppen von Nutzenden zu potenziellen Zielscheiben digitaler Gewalt werden. Politische Vorschläge befassen sich in der Regel mit Ersterem und nicht mit Letzterem. Doch beides muss gezielt angegangen werden.

Wenn sich die Content-Moderation auf stark automatisierte Entscheidungssysteme und -protokolle stützt, prallen soziale Strukturen und technische Prozesse frontal aufeinander. Insbesondere die verfassende Person eines Tweets ist Teil einer implizit geschützten Klasse, die der kommentierenden Person eines Tweets überlegen ist. (Siehe *die Dokumentation zu anstößigen Inhalten bei Twitter*.) Wenn jemand den ursprünglichen Tweet kommentiert oder darauf antwortet, erhält die kommentierende Person die Benachrichtigung, dass ihr Beitrag überprüft wurde und unter Verdacht steht, gegen die Regeln des Netzwerks zu verstoßen. Die verfassende Person des Tweets erfährt jedoch nicht die gleiche Behandlung – wie im Fall des Beitrags von Rick Santorum, der veröffentlicht blieb, während Shana Whites Tweet entfernt und ihr Account gesperrt wurde. Im Wesentlichen ist festzuhalten, dass die Sozialen Netzwerke nicht für die vorbereitende Verarbeitung und Überprüfung der Posts auf potenziell beleidigende Inhalte konzipiert wurden. Entsprechend versuchen die Plattformen grundsätzlich, Auseinandersetzungen zu unterbinden.

Eine Alternative besteht im Einsatz von Rückverfolgungs- und Verbreitungsstrategien. Die Rückverfolgung funktioniert in der Content-Moderation folgendermaßen: Wurde ein gemeldeter Post als Antwort auf einen anderen Inhalt von der Plattform entfernt, obliegt es den Protokollen der Content-Moderation, den ursprünglichen Post erneut auf die Einhaltung der Regeln zu schädlichen und illegalen Inhalte zu überprüfen. Mit anderen Worten: Wird die Antwort entfernt, wird auch der ursprüngliche Post überprüft und ggf. ebenfalls entfernt.

Die Weiterverbreitung von Inhalten funktioniert in der strategischen Content-Moderation genau umgekehrt: Wird der ursprüngliche Beitrag von der Plattform, werden auch die Antworten überprüft und ggf. entfernt. Diejenigen Personen, die den Beitrag kommentiert haben, sollten eine Mitteilung dazu erhalten, gegen welche Moderationsregeln der ursprüngliche Inhalt sowie ggf. ihr eigener Beitrag verstößt. Beim Einsatz von rückverfolgungs- oder weiterverbreitungsbezogenen Maßnahmen der Content-Moderation empfehlen wir eine konsequentere Durchsetzung. Lösungsansätze in der Content-Moderation müssen für alle Probleme desselben Typs gleichermaßen angewendet werden. Aktuell dominiert die Intransparenz der Technologie, und der Informationsfluss zur Überprüfung gepostеter Inhalte ist spärlich. Deshalb lässt sich aktuell nicht mit Sicherheit sagen, was genau passiert, wenn Beiträge in den Sozialen Netzwerken mit automatischen Entscheidungsalgorithmen geprüft werden.

Doch alle Zeichen deuten auf ein Ungleichgewicht hin, das dringend korrigiert werden muss.

Literaturverzeichnis

47 U.S. Code § 230 – Protection for private blocking and screening of offensive material. (k.A.). Cornell Law School – Legal Information Institute. *https://www.law.cornell.edu/uscode/text/47/230* (Abruf am 16.06.2021)

About. (k.A.). Shana V. White. «Illuminate others and purposefully disrupt the status quo.» Shana V. White Blog. *https://shanavwhite.com* (Abruf am 16.06.2021)

About offensive content (2021). Twitter. *https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content* (Abruf am 16.06.2021)

Angry black woman (12.06.2021). Wikipedia. The Free Encyclopedia. *https://en.wikipedia.org/wiki/Angry_black_woman* (Abruf am 16.06.2021)

Apple. (k.A.). Apple. https://www.apple.com (Abruf am 16.06.2021)

Barrett, P. M. (2020). Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern. *https://bhr.stern.nyu.edu/tech-content-moderation-june-2020* (Abruf am 16.06.2021)

Bell, T. A. (05.06.2020). It's Time We Dealt With White Supremacy in Tech. Marker. *https://marker. medium.com/its-time-we-dealt-with-white-supremacy-in-tech-8f7816fe809* (Abruf am 16.06.2021)

Bozeman, B., Johnson, J. (26.05.2014). The Political Economy of Public Values: A Case for the Public Sphere and Progressive Opportunity. SAGE Journals – The American Review of Public Administration. Band 45, Ausgabe 1. *https://journals.sagepub.com/doi/abs/10.1177/0275074014532826* (Abruf am 16.06.2021)

Brown, D'S. (05.02.2021). Male Colleagues Harass Black Female Former Googler Timnit Gebru Amid Google Ouster. Blavity News. *https://blavity.com/male-colleagues-harass-black-female-former- googler-timnit-gebru-amid-google-ouster?category1=news* (Abruf am 16.06.2021)

Buolamwini, J. (25.01.2019). Response: Racial and Gender bias in Amazon Rekognition – Commercial AI System for Analyzing Faces. Medium. *https://medium.com/@Joy.Buolamwini/response- racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced* (Abruf am 16.06.2021)

Campbell, J. @JasonSCampbell (26.04.2021). *Twitter. https://twitter.com/JasonSCampbell/status/1386685340522536961?s=20* (Abruf am 16.06.2021)

Chang, E. (05.03.2019). Brotopia. Penguin Random House. *https://www.penguinrandomhouse.com/books/547571/brotopia-by-emily-chang/* (Abruf am 16.06.2021)

Community Standard Enforcement Report (2021). Facebook. *https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement* (Abruf am 16.06.2021)

Community Standards (2021). Facebook. *https://www.facebook.com/communitystandards/introduction* (Abruf am 16.06.2021)

Deerwester, J. (k.A.). «Liam Neeson is canceled:» Fans react to actor's story of urge for racist revenge. USA Today News. *https://eu.usatoday.com/story/life/*

people/2019/02/04/liam-neeson-reveals-shocking- racially-charged-past/2766111002/ (Abruf am 16.06.2021)

Flaherty, C. (06.02.2018). Questioning «Algorithms of Oppression.» Inside Higher ED. *https://www.insidehighered.com/news/2018/02/06/scholar-sets-twitter-furor-critiquing-book-he-hasnt-read* (Abruf am 16.06.2021)

Gallo, J. A., Cho, C. Y. (27.01.2021). Social Media: Misinformation and Content Moderation Issues for Congress. Congressional Research Service. *https://crsreports.congress.gov/product/pdf/R/R46662* (Abruf am 16.06.2021)

Gorwa, R., Binns, R., Katzenbach, C. (28.02.2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. SAGE Journals – Big Data & Society. *https://journals.sagepub.com/doi/full/10.1177/2053951719897945* (Abruf am 16.06.2021)

Grimmelmann, J. (2015). The Virtues of Moderation. 17 YALE J.L. & TECH. 42 (2015). *https:// digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt* (Abruf am 16.06.2021)

Gulati-Partee, G., Potapchuk, M. (2014). Paying Attention to White Culture and Privilege: A Missing Link to Advancing Racial Equity. The Foundation Review, Band 6:1. *http://www.mpassociates.us/ uploads/3/7/1/0/37103967/paying_attention_to_white_culture_and_privilege-_a_missi.pdf* (Abruf am 16.06.2021)

Guynn, J. (k.A.). Facebook while black: Users call it getting «Zucked,» say talking about racism is censored as hate speech. USA Today News. *https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black- zucked-users-say-they-get-blocked-racism-discussion/2859593002/* (Abruf am 16.06.2021)

Hateful conduct policy (2021). Twitter. *https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy* (Abruf am 16.06.2021)

How The Facebook Ads that Targeted Voters Centered on Black American Culture: Voter Suppression was the End Game. (k.A.). SOVAW. Stop Online Violence Against Women. *https://stoponlinevaw.com/wp-content/uploads/2018/10/Black-ID-Target-by-Russia-Report-SOVAW.pdf* (Abruf am 16.06.2021)

How well do IBM, Microsoft, and Face++ AI services guess the gender of a face? (2018). Gender Shades. *http://gendershades.org* (Abruf am 16.06.2021)

Kleinman, Z. (04.02.2019). Amazon: Facial recognition bias claims are ‹misleading›. BBC News. *https://www.bbc.com/news/technology-47117299* (Abruf am 16.06.2021)

Kurenkov, A. (24.06.2020). Lessons from the PULSE Model and Discussion. The Gradient. *https://thegradient.pub/pulse-lessons/* (Abruf am 16.06.2021)

Levchak, C. C. (2018). Microaggressions and Modern Racism. Springer. *https://link.springer.com/book/10.1007/978-3-319-70332-9#about* (Abruf am 16.06.2021)

Mac, Ryan @RMac18 (k.A.). *Twitter. https://twitter.com/RMac18/status/1382366931307565057?s=20* (Abruf am 16.06.2021)

Matamoros-Fernández, A. (21.02.2017). Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Taylor & Francis Online. *https://www.tandfonline.com/doi/abs/10.1080/1369118X.2017.1293130?journalCode=rics20* (Abruf am 16.06.2021)

McClintock, E. A. (31.03.2016). The Psychology of Mansplaining. Psychology Today. *https://www. psychologytoday.com/us/blog/it-s-man-s-and-woman-s-world/201603/the-psychology-mansplaining* (Abruf am 16.06.2021)

McMillan Cottom, T. (Oktober 2019). Thick. And other Essays. The New Press. *https://thenewpress.com/books/thick* (Abruf am 16.06.2021)

Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C. (20.07.2020). PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. Duke University. *http://pulse.cs.duke.edu* (Abruf am 16.06.2021)

Messenger, H., Simmons, K. (10.05.2021). Facebook content moderators say they receive little support, despite company promises. NBC News. *https://www.nbcnews.com/business/business-news/facebook-content- moderators-say-they-receive-little-support-despite-company-n1266891* (Abruf am 16.06.2021)

Molla, R., Lightner, R. (10.04.2016). Diversity in Tech. The Wall Street Journal. *http://graphics.wsj.com/diversity-in-tech-companies/* (Abruf am 16.06.2021)

Núñez, A-M., Mayhew, M. J., Shaheen, M., Dahl, L. S. (15.03.2021). Let's Teach Computer Science Majors to Be Good Citizens. The Whole World Depends on It. EdSurge. *https://www.edsurge.com/ news/2021-03-15-let-s-teach-computer-science-majors-to-be-good-citizens-the-whole-world-depends-on-it* (Abruf am 16.06.2021)

Our range of enforcement options (2021). Twitter. *https://help.twitter.com/en/rules-and-policies/enforcement-options* (Abruf am 16.06.2021)

Rooney, K., Khorram, Y. (12.06.2020). Tech companies say they value diversity, but reports show little change in last six years. CNBC. *https://www.cnbc.com/2020/06/12/six-years-into-diversity-reports-big-tech-has-made-little-progress.html* (Abruf am 16.06.2021)

Schiffer, Z. (05.03.2021). Timnit Gebru was fired from Google – then the harassers arrived. The Verge. *https://www.theverge.com/platform/amp/22309962/timnit-gebru-google-harassment-campaign-jeff-dean* (Abruf am 16.06.2021)

Shana V. White @ShanaVWhite (2021). Twitter. *https://twitter.com/ShanaVWhite/status/1397873437197078530?s=20* (Abruf am 16.06.2021)

Shea Wesley Martin @sheathescholar (2021). Twitter. *https://twitter.com/sheathescholar/status/1386744704176431104?s=20* (Abruf am 16.06.2021)

The Twitter Rules (2021). *Twitter. https://help.twitter.com/en/rules-and-policies/twitter-rules* (Abruf am 16.06.2021)

The State of Black & Latinx Women Founders (2021). Digitalundivided. *https://www.projectdiane.com* (Abruf am 16.06.2021)

They aren't talking about me… (14.03.2010). The Crunk Feminist Collection. *http://www.crunkfeministcollective.com/2010/03/14/they-arent-talking-about-me/* (Abruf am 16.06.2021)

Turner, K., Wood, D., D'Ignazio, C. (27.01.2021). The Abuse and Misogynoir Playbook. mit media lab. *https://www.media.mit.edu/articles/danielle-wood-and-katlyn-turner-co-author-article-the-abuse-and- misogynoir-playbook-for/* (Abruf am 16.06.2021)

Vogels, E. A. (13.01.2021). The State of Online Harassment. Pew Research Center. *https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/* (Abruf am 16.06.2021)

Wood, M. (26.01.2019). Thoughts on Recent Research Paper and Associated Article on Amazon Rekognition. AWS Machine Learning Blog. *https://aws.amazon.com/de/blogs/machine-learning/thoughts- on-recent-research-paper-and-associated-article-on-amazon-rekognition/* (Abruf am 16.06.2021)

Die Autorin

Brandeis Marshall ist Professorin für Computerwissenschaften am Spelman College sowie Practitioner Fellow am Stanford University Digital Civil Society Lab. Sie lehrt und forscht in den Bereichen Datentechnik, Datenwissenschaften und Informatik. Ihre Schwerpunktthemen sind die Auswirkungen von Daten in der Informationstechnologie mit Hinblick auf ethnische Zugehörigkeit, Geschlecht und sozioökonomischen Status. Darüber hinaus ist sie Gründerin des Schulungsunternehmens DataedX, das Führungskräfte bei der Optimierung der Datenkompetenzen und der Karriereentwicklung ihrer Teams unterstützt.

Impressum

Herausgeberin: Heinrich-Böll-Stiftung e.V., Schumannstraße 8, 10117 Berlin
Kontakt: Gunda-Werner-Institut für Feminismus & Geschlechterdemokratie,
Katharina Klappheck **E** *klappheck@boell.de*

Erscheinungsort: www.boell.de
Erscheinungsdatum: Oktober 2022
Covermotiv: © Pia Danner
Übersetzung: Translationes

Weitere E-Paper zum Downloaden unter
www.boell.de/publikationen