Schattauer

Von Hass-Sprache zur Kriminalität

Facebook, Twitter, die AfD und der Brexit

Korrespondenzadresse
Prof. Dr. Dr. Manfred Spitzer
Universität Ulm
Abteilung für Psychiatrie
Leimgrubenweg 12–14
87054 Ulm

Bibliografie
DOI https://doi.org/10.1055/a-0952-6895
Nervenheilkunde 2019; 38: 794–802

ISSN 0722-1541

ZUSAMMENFASSUNG

Soziale Online-Medien verbreiten Sprache und damit auch Hass-Sprache. Sie wirken damit verstärkend auf die Auswirkungen von Hass-Sprache und befeuern Hass-Kriminalität, wie anhand zweier Studien zu den Auswirkungen von Facebook in Deutschland und von Twitter in Großbritannien gezeigt wird. Hass-Sprache als Ausdruck der freien Rede (wie in den USA) zu schützen, erscheint im Hinblick auf diese empirischen Erkenntnisse weniger geboten als behutsame Zensur, begleitet von Vernunft und dem beherzten Eintreten für Freiheit. Durch die heute mögliche Erfassung von Hass-Sprache in Echtzeit wird eine neue Form der prädiktiven Polizeiarbeit möglich, deren Vor- und Nachteile sorgfältig abzuwägen sind.

Dass Worte zu Taten werden können, gehört zu ihren bekannten Wirkungen. Dass technische Neuerungen im Bereich der Kommunikation die Gefahr beherbergen, diesen Mechanismus zu verstärken, wissen wir spätestens seit den Erfindungen des Buchdrucks und des Volksempfängers. Gegenwärtig stehen vor allem die negativen sozialen Auswirkungen von sozialen Online-Medien wie Facebook oder Twitter im Fokus des Interesses [14–16]. In letzter Zeit machten vor allem Hass-Kommentare gegen Immigranten und der Einwanderung überhaupt auf Facebook und Twitter die Runde, sodass diese Firmen aufgefordert wurden, hier eine bessere „Zensur“ einzuführen. Ist das sinnvoll oder gar notwendig? Kann die Sozialwissenschaft bei dieser Frage die Politik beraten?

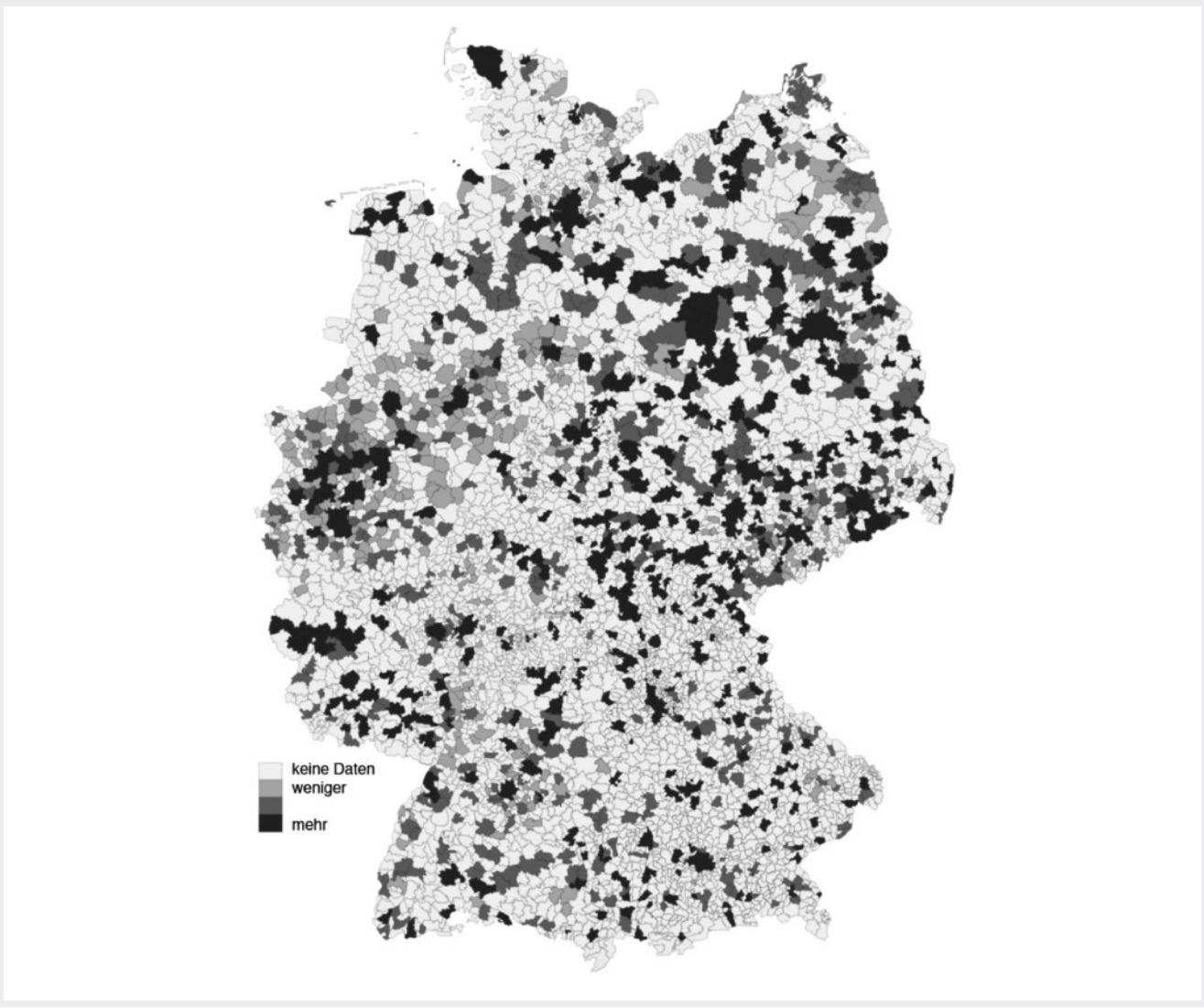
Hass in sozialen Online-Medien

Als Hintergrundinformation sollte man hierzu wissen, dass Facebook in den USA von mehr als der Hälfte der Bevölkerung als Nachrichtenquelle (also wie hierzulande die „Tagesschau“) genutzt wird [12] und dass dies hierzulande für die 18- bis 25-Jährigen ebenfalls gilt, wenn auch in geringerem Ausmaß [4]. Da man sich in Facebook die Nachrichten gewissermaßen aussuchen kann, führt dies dazu, dass man genau diejenigen Fakten und Meinungen mitgeteilt bekommt, die man lesen möchte. Man nennt dies den Echokammer-Effekt, der die besondere Leichtigkeit der thematischen Einengung der Bevölkerung in großem Stil durch soziale Medien wie Facebook beschreibt.

Die Leichtigkeit und Schnelligkeit der Verwendung sozialer Online-Medien wirkt sich auf die Qualität der von ihnen verbreiteten Statements aus: Schreibt man etwas auf Papier auf, lässt man es vielleicht liegen, denkt noch einmal darüber nach und überdenkt vielleicht auch manche im Affekt getätigte Äußerung. Erscheint etwas in gedruckter Form kommt ein kritischer Copy-Editor und vielleicht noch ein Chefredakteur oder Herausgeber hinzu, was Qualität und Objektivität der Nachricht weiter verbessern kann. Soziale Online-Medien hingegen verleiten dazu, Gedanken ungefiltert durch den eigenen oder fremden kritischen Verstand einfach in die Welt zu posaunen, und dies wiederum vermindert die Wahrhaftigkeit und vermehrt den Emotionsgehalt der öffentlich gemachten Äußerungen. Manche sehen das als „Gewinn an Authentizität“. Andere als Verlust von Wahrheit und Vernunft. Was ist nun tatsächlich der Fall?

Man könnte argumentieren, dass solche verbalen Attacken nur ein Ventil darstellen, durch das manche Menschen „Dampf ablassen“, sodass eine Minderung des „Drucks“ und damit ein positiver Gesamteffekt resultiert. Hass-Kommentare hätten demnach – im Sinne einer Art von Katharsis – keine oder eine mindernde Auswirkung auf entsprechende Handlungen, d. h. auf Hass-Kriminalität. Hass-Sprache fällt nach der US-amerikanischen Verfassung unter die Freiheit der Rede und kann dort daher auch nicht verboten werden.

Hierzulande (und allgemein in der EU) wird dies anders gesehen. Die Freiheit des einen hört dort auf, wo die Freiheit des anderen beeinträchtigt wird, was angesichts der – aus Interviews und entsprechender qualitativer Forschung bereits bekannten – negativen Auswirkungen von Hass-Sprache auf die Betroffen nachzuvollziehen ist. Man kann zudem argumentieren, dass auf Reden Taten folgen können, dass der verbale Hass also aggressive, kriminelle Handlungen bahnen und sie damit überhaupt erst hervorbringen könnte. In Deutschland ist seit dem 1.10.2017 das Netzwerkdurchsetzungsgesetz (Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken – NetzDG), auch „Facebook-Gesetz“ genannt, in Kraft. Es regelt den Umgang mit Nutzer-Beschwerden über Hasskriminalität und andere strafbare Inhalte in sozialen Netzwerken wie Facebook, YouTube und Twitter und führte Bußgelder für diese Fir-



▶ **Abb. 1** Verteilung der 3335 untersuchten Gewaltdelikte (bezogen auf die Anzahl von Asylsuchenden) in den 4466 untersuchten Gemeinden (nach Daten aus [10]. Die Daten zu den Gemeinden entstammen einer online zugänglichen Datenbank des Gemeindeverwaltungsverbandes. Daher ist die Zahl der Gemeinden hier geringer als die Anzahl aller Dörfer und Städte in Deutschland, die am 31.12.2017 mit 11 054 [17] mehr als doppelt so hoch lag. Man verwendete jedoch die genannten Daten, weil sie besser zugänglich und im Hinblick auf die Einwohnerzahlen homogener waren (In der Datenbank des Gemeindeverwaltungsverbandes werden kleine Gemeinden nicht getrennt betrachtet).

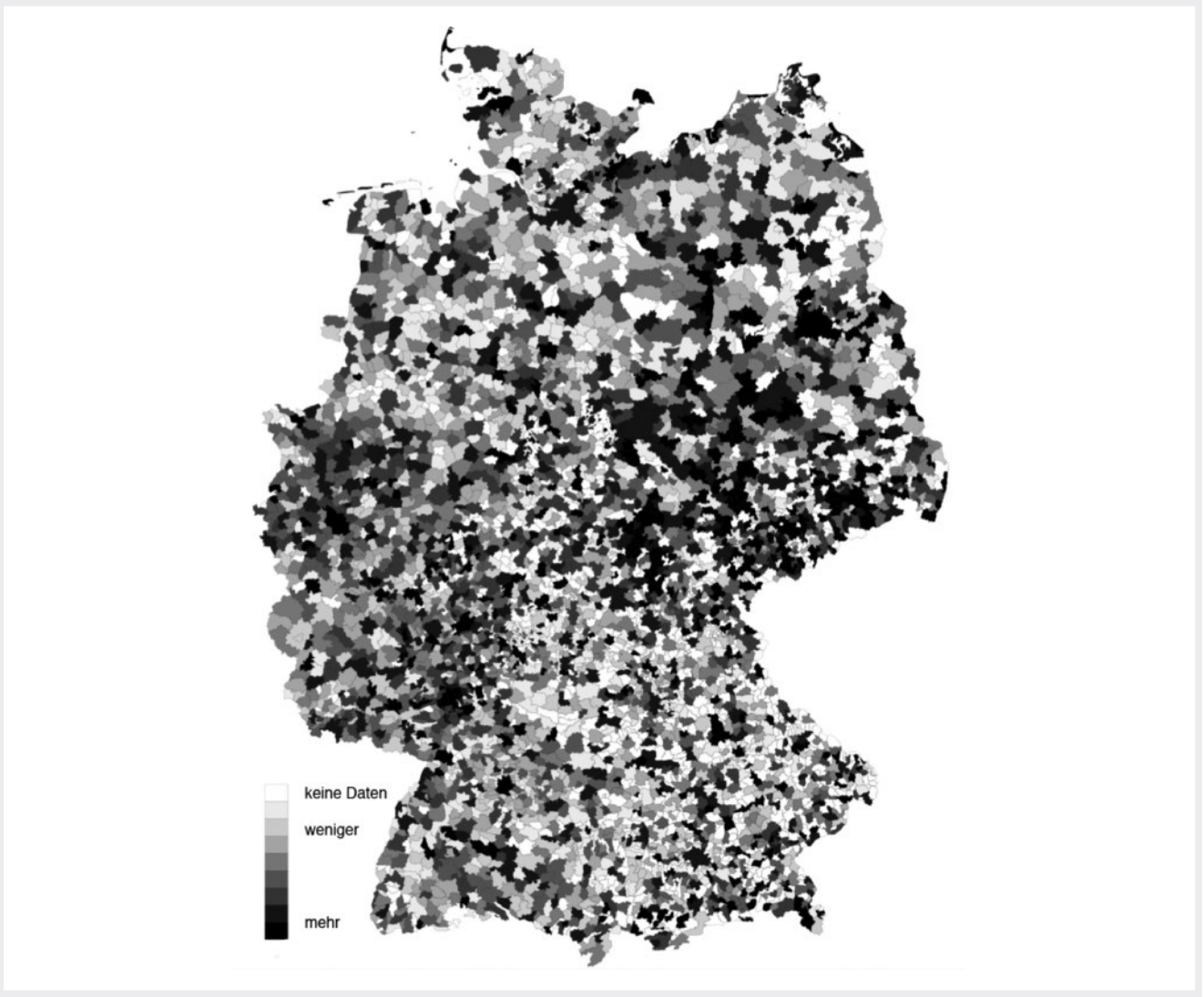
men ein, wenn sie Hass-Sprache verbreiten. Mit dem Gesetz werden Anbieter von sozialen Netzwerken dazu verpflichtet, rechtswidrige Inhalte im Sinne des NetzDG nach Kenntnis und Prüfung zu entfernen oder den Zugang zu ihnen zu sperren. Ist dies sinnvoll, ist es richtig? – Der Frage versuchen 2 Studien, eine über Deutschland und die AfD und eine zweite über Großbritannien und den Brexit, empirisch nachzugehen.

Facebook und die AfD

Karsten Müller (Princeton University, USA) und Carlo Schwarz (University of Warwick, GB) untersuchten schon vor einem Jahr anhand von Facebook-Daten den Zusammenhang zwischen Hass-Kommentaren auf sozialen Online-Medien und Hass-motivierter Kriminalität in Deutschland. Sie nutzten die historische Tatsache, dass in den Jahren 2015 und 2016 etwa eine Million Flüchtlinge nach Deutschland kamen und dass dies eine Welle krimineller Handlungen gegenüber Flüchtlingen nach sich zog. Sie nutzten zudem die Tatsache, dass die Partei Alternative für Deutschland (AfD) – bekannt für ihre abwertende Haltung gegenüber Flüchtlingen und Einwanderung – mit 420 000 „Followern“ von allen deutschen Parteien die größte Gefolgschaft in Facebook hat.

Als erstes wurde unter Bezugnahme auf Daten der Stiftung Amadeu Antonio sowie der Nichtregierungsorganisation Pro Asyl (sowie entsprechenden Pressemitteilungen) 3335 Gewaltdelikte gegenüber Flüchtlingen in Deutschland im Zeitraum von 2015 bis 2017 ermittelt, darunter 534 Fälle von Körperverletzung und 225 Fälle von Brandstiftung (▶ **Abb. 1**).

Analysiert wurde zum zweiten eine große Zahl von Daten zu Hass-Statements gegenüber Flüchtlingen auf der Facebook-Seite der AfD: 176 153 Meinungsäußerungen („Facebook-posts“), 290 854 Kommentare und 510 268 Zustimmungen („Likes“). Diese



▶ **Abb. 2** Anzahl der Nutzer der Facebook-Seite der AfD (bezogen auf die Anzahl der Einwohner) in den 4466 untersuchten Gemeinden (nach Daten aus [10]).

wurden von insgesamt 93 806 Personen im Untersuchungszeitraum abgegeben. Man muss dazu wissen, dass die Facebook-Seite der AfD die Besonderheit hat, dass Parteimitglieder auf ihr direkt Meinungen äußern können (was auf den Seiten anderer Parteien so nicht geht). Dies erlaubte es den Autoren, eine Art Messgröße für das deutschlandweite wöchentliche Ausmaß von flüchtlingsfeindlichen verbalen Äußerungen in Facebook – die Autoren sprechen von „measure for anti-refugee hate speech salience on social media“ [10] – zu berechnen.[1]

Zum Dritten berechneten die Autoren für jede Gemeinde einzeln das Ausmaß der wöchentlichen Facebook-Nutzung. Da diese Daten nicht vorliegen (Daten zur Facebook-Nutzung gibt es nur für Gesamtdeutschland sowie auf der Ebene der 16 Bundesländer) musste dies „von Hand“ durchgeführt werden. Von den 93 806 Nutzern der AfD Facebook-Seite konnte der Ort in 39 632 Fällen ermittelt werden (▶ **Abb. 2**).

Die Logik dahinter beschreiben die Autoren wie folgt: „Wir postulieren, dass die Informationskanäle der sozialen Online-Medien flüchtlingsfeindliche Ressentiments verstärken können, was manche potenziellen Gewalttäter zu Gewaltakten verleiten könnte. Wenn soziale Online-Medien hierbei eine Rolle spielen, würden wir erwarten, dass mehr Hass-motivierte kriminelle Delikte in Gemeinden mit einer höheren Facebook-Nutzung auftreten, insbesondere in Zeiten erhöhter Spannung und Konfliktbereitschaft“ [10].[2]

Die Autoren analysierten also Daten aus 4 466 deutschen Gemeinden über einen Zeitraum von 111 Wochen (1.1.2015–

1 Wer (noch) mehr Details zum Vorgehen der Autoren wissen möchte, dem sei das Studium des sehr aufschlussreichen und sehr gut geschriebenen Methoden-Teils der Arbeit empfohlen.

2 „We posit that social media information channels can reinforce anti-refugee sentiments, which may push some potential perpetrators over the edge to carry out violent acts. If social media plays a role, we would expect more hate crimes to occur in municipalities with higher exposure to Facebook, particularly when tensions are high.“

13.2.2017) zum Auftreten von den 3335 Delikten gegenüber Flüchtlingen. Sie konnten mit Hilfe dieser 3 gemessenen Variablen zeigen, dass die Präsenz von negativen Aussagen über Flüchtlinge auf der AfD-Facebook-Seite mit anschließenden kriminellen Akten gegenüber Flüchtlingen in Zusammenhang steht, und dass dieser Zusammenhang umso größer ist, je stärker Facebook genutzt wird. Korrelation ist nicht dasselbe wie Kausalität. Denn es könnte ja sein, dass der gefundene Zusammenhang zwischen Facebook-Nutzung und Hass-Kriminalität auf Gemeindeebene durch andere unbeobachtbare oder beobachtete Faktoren zustande kommt, d. h. nicht durch die Facebook-Nutzung verursacht ist.

Um hier weiter zu kommen und ein Ursache-Wirkungs-Verhältnis zwischen Facebook-Nutzung und Hass-Kriminalität nachzuweisen, verwendeten die Autoren das, was man ein quasi-experimentelles Design nennt, und machten sich die Tatsache zu Nutze, dass erstens das Internet zuweilen lokal ausfällt und es zweitens bei Facebook manchmal landesweit zu Ausfällen kommt. Man suchte also nach Internetausfällen in Gemeinden, jeweils auf eine bestimmte Woche bezogen. Diese ließen sich anhand von Beschwerden von Nutzern finden, was wiederum durch entsprechende lokale und nationale Pressemeldungen über diese Ausfälle verifiziert wurde. Diese lokalen Ausfälle führen zu einer geringeren lokalen „Berieselung" der Bevölkerung durch Hass-Kommentare in den sozialen Medien, wohingegen das deutschlandweite „Ausgesetztsein" sich nicht ändert. „Es sei hervorgehoben, dass Internet-Ausfälle geografisch überall auftreten und weder mit den AfD-Likes auf Facebook noch mit der Anzahl flüchtlingsfeindlicher Delikte von Hass-Kriminalität in Zusammenhang stehen" kommentieren die Autoren ihr Vorgehen [9].[3]

Mit diesem Verfahren wurde nachgewiesen, dass die Steigerung der Hass-Kriminalität in Zeiten von mehr Hass-Kommentaren auf Facebook an Orten und Zeiten von Internet-Ausfällen völlig ausblieb. Man konnte weiterhin zeigen, dass der Effekt auf die Facebook-Nutzung und nicht auf die Internet-Nutzung im Allgemeinen zurückging.[4] Für einen kausalen Effekt von Facebook auf die Auswirkungen von Hass-Kommentaren im Sinne der Befeuerung von Hass-Kriminalität spricht weiterhin der Befund, dass landesweite Facebook-Ausfälle zu einem Verschwinden des Zusammenhangs zwischen Hass-Kommentaren auf Facebook und tatsächlicher Hass-Kriminalität führten.[5]

Als wäre dies noch nicht genug der Evidenz, führen die Autoren des Weiteren vor, was heutzutage mit den Mitteln der digitalen Welt einerseits und der empirischen Sozialforschung andererseits möglich ist. Während sich in den bisherigen Analysen das Ausmaß der Facebook-Nutzung auf AfD-Unterstützer bezog, wurde für eine weitere Analyse ein Maß für die Facebook-Nutzung überhaupt verwendet, unabhängig von der Mitgliedschaft in der AfD. Hierzu benutzten die Autoren Daten der Facebook-Seite von Nutella, die mit über 32 Millionen Likes zu den beliebtesten Facebook-Seiten Deutschlands gehört und damit Aussagen über die Facebook-Nutzung durch die allgemeine Bevölkerung erlaubt. Dieser Schokoladen-Brotaufstrich wird „von Menschen jeglichen sozioökonomischen Hintergrunds und in allen Regionen [Deutschlands] konsumiert", kommentieren die Autoren und bemerken zudem, dass „die breite Beliebtheit [von Nutella] dessen Facebook-Seite für unsere Analysen attraktiver macht als andere populäre Facebook-Seiten wie die des FC Bayern München oder die von BMW, deren Nutzer sich möglicherweise um eine bestimmte geografische Gegend oder sozioökonomische Schicht gruppieren" [9].

Durch diese Analyse konnte gezeigt werden, dass in Gemeinden mit vielen Nutzern der Nutella-Facebook-Seite pro Kopf bei starkem Vorhandensein von Hass-Reden (auf der AfD-Facebook-Seite) ebenfalls die Anzahl der Hass-Delikte anstieg – es sei denn, die Facebook-Nutzung war gerade nicht möglich, weil das Internet lokal oder Facebook deutschlandweit nicht funktionierte. Es ist also die Facebook-Nutzung ganz allgemein (und nicht allein die Nutzung der AfD-Facebook-Seite), die sich als Mediator des Zusammenhangs von verbalem Hass und realer Gewaltkriminalität erweist.

Es sei noch einmal betont, dass es hier nicht darum ging, zu zeigen, dass Facebook flüchtlingsfeindliche Inhalte produzieren würde. Vielmehr ging es darum, dass Facebook offensichtlich wesentlich beteiligt ist an der Verbreitung flüchtlingsfeindlicher Inhalte als Grundlage von Hass-Kriminalität, woher diese Inhalte auch kommen und wie auch immer sie entstehen mögen.

Twitter: Mehr Hass, weniger Wahrheit

Ähnliches gilt auch für Twitter, wie die gleichen Autoren in einer anderen Studie mit dem ursprünglich sehr pointierten Titel Making America Hate Again? Twitter and Hate Crime Under Trump fanden[6] [10]. Für Twitter wurde ebenfalls im Jahr 2018 gefunden, dass dieses Medium Falschheit rascher, weiter und tiefer verbreitet als Wahrheit, ohne dass es die Falschheit selbst produziert [18] (▶ **Abb. 3**). Es stellt jedoch gleichsam einen Turbo für die Verbreitung von Gerüchten dar.

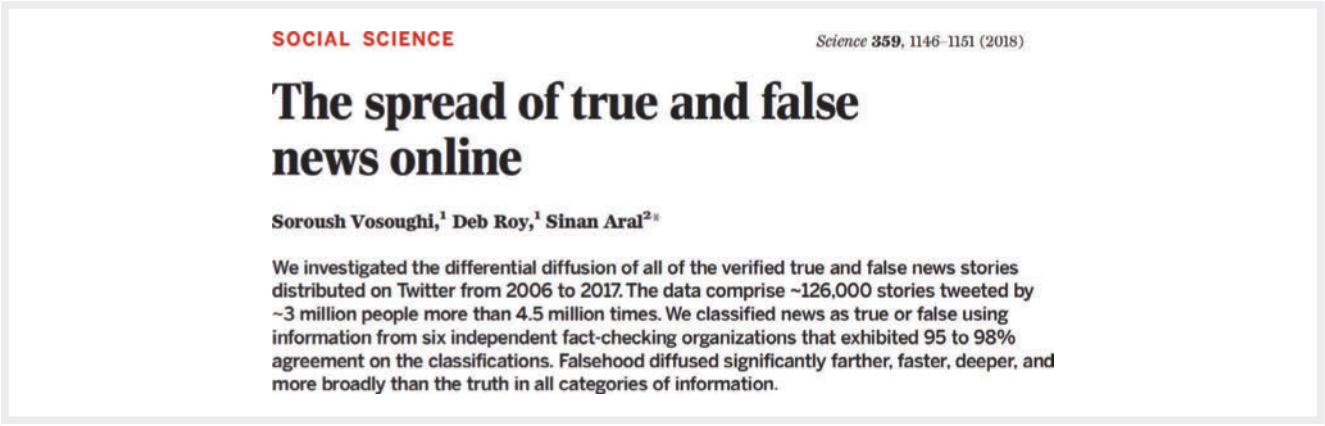
Eine kürzlich publizierte Studie zu mittels Twitter verbreiteten rassistischen und religiösen Hass-Aussagen („Hate Speech") in Großbritannien bestätigt im Wesentlichen die für Deutschland und Facebook gefundenen Ergebnisse [19]. Matthew Williams und Mit-

3 „Notably, internet disruptions are geographically dispersed and orthogonal to both AfD likes on Facebook and the total number of refugee attacks in a municipality."

4 Ausnahmsweise sei hier der gesamte Absatz im Original wiedergegeben, um den methodischen Detailreichtum der Arbeit zu illustrieren: „We find that, while the number of hate crimes increases during periods of higher anti-refugee salience, this effect disappears for municipalities experiencing an internet outage. Quantitatively, a typical internet disruption fully mediates social media's effect on hate crime. Further, internet outages themselves do not affect the number of anti-refugee incidents beyond their impact through social media usage. This makes it unlikely that we are capturing a ‚displacement effect' which could arise if potential perpetrators were merely busy fixing their internet access. This further points to social media as the propagation mechanism. Additionally, we do not appear to be capturing other online channels: internet outages have no mediating effect on hate crimes in areas with higher general internet usage, once we take social media usage into account" [10].

5 Im Original: „Consistent with a causal effect of social media, we again find that the effect of anti-refugee salience on hate crimes essentially disappears in weeks of major Facebook outages" [9].

6 Der Titel der Arbeit wurde mittlerweile in From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment geändert.



SOCIAL SCIENCE

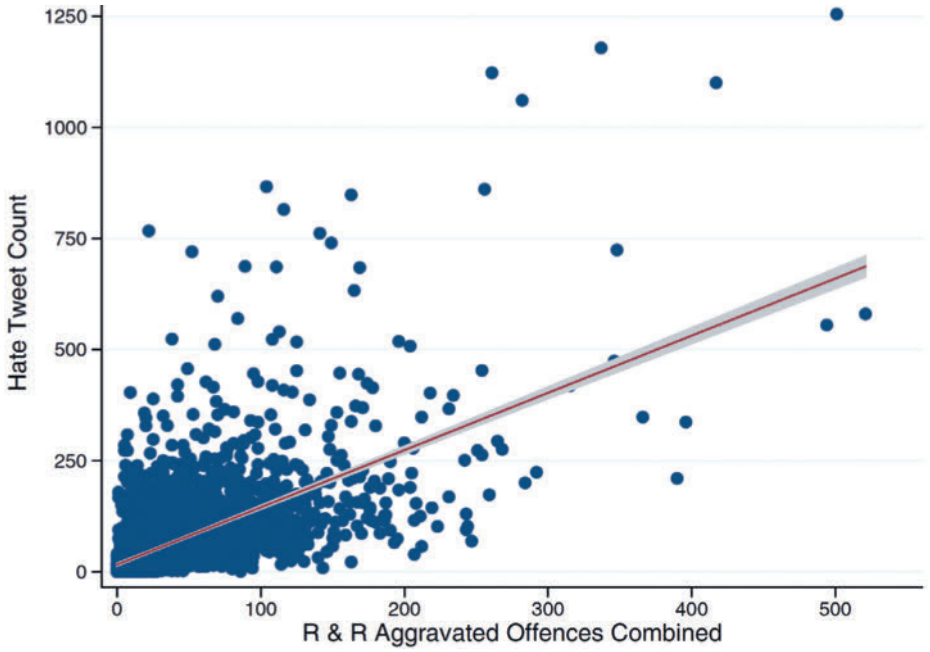
Science **359**, 1146–1151 (2018)

The spread of true and false news online

Soroush Vosoughi,[1] Deb Roy,[1] Sinan Aral[2]*

We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information.

▶ **Abb. 3** Facsimile des Beginns der Arbeit von Vosoughi und Mitarbeitern [18] zum Gerüchte-Turbo Twitter.



▶ **Abb. 4** Zusammenhang zwischen Hass-Sprache (Y-Achse: Anzahl von Hass-Statements auf Twitter) und Hass-Kriminalität (X-Achse: rassistische und religiöse Hass-motivierte Verstöße (R & R aggravated offences, nach Daten aus [18], ▶ Fig. 7).

arbeiter von der Cardiff University stellen in ihrer Arbeit zunächst fest, dass in England und Wales in den Jahren 2017/18 mit 94 098 Hass-Delikten ein historischer Höchststand von Hass-Kriminalität erreicht worden sei. Die Steigerung gegenüber dem Vorjahr betrug 17 %, die gegenüber 2012/13 betrug 123 %, wobei die Hass-Kriminalität mit rassistischem („anti-black“) oder religiösen („anti-muslim“) Hintergrund besonders stark angestiegen sei, und eine nicht unbeträchtliche Dunkelziffer noch hinzukäme.

Die Autoren sammelten Daten zu Hass-Kriminalität und Twitter Hass-Statements in London über einen Zeitraum von 8 Monaten in den Jahren 2013/14. Insgesamt 165 990 religiös oder rassistisch motivierte Vergehen wurden identifiziert, dazu 21,7 Millionen in 4720 Londoner Bezirken lokalisierten Twitter-Nachrichten („posts“), die mittels eines durch machine-learning trainierten Klassifizierungs-Netzwerks in Hass-posts und Nicht-Hass-posts eingeteilt wurden. Dies resultierte in einem Datensatz von 294 361 Twitternachrichten (1,4 % aller 21,7 Millionen Nachrichten) im gleichen Zeitraum. Alle Daten waren auf den Monat und den genauen Ort bezogen.[7] Aus Census-Daten wurden weiterhin folgende Variablen mit der gleichen Zeit- und Ortsauflösung ermittelt:

- Anteil der Menschen ohne Ausbildung,
- Anteil der 14- bis 24-Jährigen,
- Anteil der Langzeitarbeitslosen und
- Anteil der Menschen aus ethnischen Minderheiten.

7 Für die Ortskodierung wurde das LOSA System (Lower Layer Super Output Area) verwendet, eine in der britischen Sozialwissenschaft häufig benutzte Datenquelle.

Schattauer

Durch komplexe statistische Analysen dieser Daten konnten Williams und Mitarbeiter einen deutlichen Zusammenhang zwischen Hass-Reden auf Twitter und religiöser oder rassistischer Hass-Kriminalität nachweisen (▶ **Abb. 4**). „Über einen 8-monatigen Beobachtungszeitraum wurde gefunden, dass auf Rasse und Religion abzielende Online-Hass-Sprache in einem Zusammenhang mit rassistisch und religiös verstärkter Offline-Hass-Kriminalität steht, einschließlich der gesamten Hass-Kriminalität überhaupt. Die Größe des Effekts ist für Delikte verschiedener Kategorien relativ gleich“,[8] schreiben Williams und Mitarbeiter hierzu [19].

Einfluss auf die Hass-Kriminalität im Sinne einer Steigerung hatten erstens das Fehlen einer Ausbildung und zweitens Langzeitarbeitslosigkeit. Der relative Anteil von Minderheiten in der Bevölkerung einer Gegend hängt umgekehrt-U-förmig mit Gewalt zusammen: Ein geringer und ein sehr großer Minderheitenanteil in der Bevölkerung bewirkt eher weniger Kriminalität. Überdeutlich wurde damit nicht nur, dass Hass-Kriminalität und Hass-Sprache in Zusammenhang stehen, sondern auch, dass diese Effekte lokal ganz unterschiedlich verteilt sind und zum Teil mit lokalen Besonderheiten interagieren.

Insgesamt ergibt sich zusammen mit der zuvor diskutierten Arbeit über die Verhältnisse in Deutschland: Der Beitrag von sozialen Online-Medien wie Facebook und Twitter zur Verbreitung von Hass und damit zu tatsächlicher Gewaltkriminalität im Sinne einer kausalen Beziehung ist sehr wahrscheinlich. Was folgt?

Predicitve Policing

Im US-amerikanischen Science-Fiction-Thriller Minority Report[9] des Regisseurs Steven Spielberg mit Tom Cruise in der Hauptrolle aus dem Jahr 2002 werden im Washington des Jahres 2054 kriminelle Delikte vorhergesagt und dadurch verhindert. Was Science Fiction war, ist mittlerweile Wirklichkeit: Predictive Policing (auf Deutsch sperrig: vorhersagende Polizeiarbeit), d. h. die Analyse von Daten aus einzelnen Fällen zur Berechnung der Wahrscheinlichkeit zukünftiger Straftaten mit dem Ziel, die Arbeit der Polizei effektiver zu machen, gibt es seit einigen Jahren und wird – meist in zeitlich befristeten Test- oder Pilotbetrieben – in Deutschland (nur ortsbezogen, z. B. zur Prävention von Wohnungseinbrüchen) in verschiedenen Bundesländern sowie in der Schweiz, Großbritannien und den USA (auch personenbezogen) eingesetzt – mit eher bescheidenen Resultaten [2, 6].

Gerade die personenbezogene Vorhersage ist jedoch problematisch, da die hierfür verwendeten lernenden Maschinen auch menschliche Vorurteile lernen oder sogar verstärken können. So wurden beispielsweise in New York, wo Predictive Policing in den 1990er-Jahren erstmals zur Vorbeugung gegen die überbordende Gewaltkriminalität zum Einsatz kam, vor allem bei der schwarzen Bevölkerung Rauschgift und Waffen gefunden. Dies lag jedoch u. a. daran, dass vor allem dieser Teil der Bevölkerung kontrolliert wurde (also ein Sampling Bias bestand). Aus diesen Daten wurde vorschnell geschlossen, das man vor allem die schwarze Bevölkerung kontrollieren müsse, was die Wahrscheinlichkeit, bei dieser Bevölkerungsgruppe fündig zu werden (und bei der weißen Bevölkerungsgruppe nicht!) noch weiter erhöhte. Derartige Schwierigkeiten begleiten das Predictive Policing bis heute, weswegen vielfach Kritik an dieser Praxis geübt wurde [5, 6]. Der indische Menschenrechtsaktivist Salil Shetty, von 2010 bis 2018 Generalsekretär der Menschenrechtsorganisation Amnesty International, äußerte sich bereits im Jahr 2016 kritisch [13]:

„Predictive Policing macht einen signifikanten Schritt [dahingehend], dass Leute oder Gruppen identifiziert werden, von denen angenommen wird, dass sie eine Straftat begehen könnten, bevor diese selbst den Gedanken daran haben, dies zu tun. Basierend auf vorhandenem Wissen über vergangene Delikte verwendet [vorhersagende Polizeiarbeit] künstliche Intelligenz, um die Wahrscheinlichkeit und den Ort von Straftaten vorherzusagen, bevor sie geschehen. Aber es gab schon jede Menge Kritik, dass dieses Vorgehen bereits bestehende Vorurteile gegenüber religiösen oder ethnischen Minderheiten verstärkt. […] Auf einer noch grundlegenderen Ebene gilt: Wenn man Leute schon zu einem Zeitpunkt als Kriminelle behandelt, bevor sie überhaupt die Absicht hatten, eine Straftat zu begehen, werden unsere Begriffe von Unschuld und freiem Willen vollkommen untergraben“[10] [13].

Da man – wie die beiden ausführlich dargestellten Studien für Deutschland und Großbritannien zeigen – Hass-Sprache in Online-Medien permanent, d. h. in Echtzeit, erfassen kann, liegt die Idee nahe, den Zusammenhang von Hass-Sprache und Hass-Kriminalität zur Prävention von Hass-Kriminalität einzusetzen. Es handelt sich hierbei um eine neue Form des Predictive Policing, die nicht auf Daten der Polizei zu kriminellen Akten, sondern auf Daten zu Mitteilungen in sozialen Online-Medien zurückgreift. Hierzu werden ebenfalls die Methoden des Machine-Learning aus dem Bereich der Artificial Intelligence (AI) verwendet. In einer kleinen, im britischen Fachblatt New Scientist am 28.8.2019 publizierten Übersicht hierzu [8] wird die Entwicklung eines Computersystems beschrieben, das mit Hilfe der in der Studie von Williams und Mitarbeitern gewonnenen Erkenntnisse unter deren Mitwirkung entwickelt wurde.

Online-Hass-Sprache und der Brexit

Vor dem Hintergrund, dass sich in Großbritannien gerade historische Ereignisse abspielen, die zu einer spürbaren Zunahme von Emotionalität, Radikalität, politischer Instabilität und Hass-Sprache geführt haben, überwacht die britische Polizei mittlerweile täglich Hunderttausende Brexit-bezogene Nachrichten (Tweets)

8 „Online hate speech targeting race and religion is positively associated with all offline racially and religiously aggravated offences, including total hate crimes in London over an eight-month period. The magnitude of the effect is relatively even across offence category“, kommentieren die Autoren ihre Ergebnisse [19].

9 Der Film basiert auf der Kurzgeschichte des US-amerikanischen Science-Fiction-Autors Philip K. Dick. Sie erschien zunächst in der Zeitschrift Fantastic Universe im Jahr 1956.

10 „Predictive policing takes a very significant step […] to identify people or groups who it believes could commit a crime before they have any actual intention of doing it. Based on existing intelligence about past crimes it uses artificial intelligence to identify the likelihood and location of crimes before they occur. But there has already been a lot of criticism for reinforcing existing biases against racial and religious minorities. […] And even more fundamentally, if people are treated as criminals when they have not even had the intent to commit crime, our notions of innocence and guilt are completely subverted” [13].

auf Twitter. Dies geschieht mittels eines „Online Hate Speech Dashboard“, das von Mitarbeitern einer vom britischen Innenministerium (Home Office) im Jahr 2017 einberufenen Institution, dem „National Police Chiefs’ Council’s online hate crime hub“, entwickelt wurde. Diese „Instrumententafel für Online Hass-Sprache“ ermöglicht es, irgendwo im Land verstärkt auftretende Hass-Sprache (islamophob, antisemitisch, Minderheiten-diskriminierend etc.) zu entdecken, während sie gerade entsteht, um falls nötig Gegenmaßnahmen zu treffen. „Das Dashboard zeigt täglich zwischen 500 000 und 800 000 Tweets zum Brexit an, von denen 0,2 bis 0,5 % als „hasserfüllt“ klassifiziert werden. Von diesen können 0,2 % [d. h. 200 bis 320 Tweets] innerhalb von Städten in Großbritannien geortet und vom Dashboard auf einer Landkarte als Hass-Hotspots angezeigt werden. Wenn irgendwo ein Spitzenwert auftaucht, kann diese Information von den Analysten an die relevante lokale Polizei weitergegeben werden,“[11] wie im New Scientist beschrieben wird [8].

Die detaillierten Informationen zu Hass gegen verschiedene Gruppen und dessen Verlauf über die Zeit an bestimmten Orten sollen helfen, Online-Hass-Sprache zu vermindern. Und obwohl die Daten anonym sind, kann man Netzwerke und koordinierte Hate-Speech-Attacken beobachten und womöglich einschreiten, bevor etwas überkocht. Einzelne kriminelle Akte wird man also eher nicht verhindern können, Aufstände und bürgerkriegsartige Zustände aber vielleicht. Weil nach dem Brexit-Votum in Großbritannien im Juli 2016 die Hass-Reden deutlich zugenommen hatten (gegenüber dem Juli im Vorjahr um 44 %), geht man derzeit davon aus, dass dies in den Wochen vor dem wahrscheinlichen tatsächlichen Austritt Ende Oktober 2019 erneut geschehen könnte. Und genau deswegen kommen die genannten Studien zur richtigen Zeit. Vielleicht kann mit ihrer Hilfe ja tatsächlich reale Gewalt verhindert werden.

Maschinelle Vorurteile und die Freiheit der Rede

Kritisch ist allerdings anzumerken, dass lernende Maschinen auch aus Sprachinput Vorurteile generieren können. So wurde beispielsweise nachgewiesen, dass allein das Vorkommen von Eigenheiten der Sprache der schwarzen Bevölkerung in den USA die Wahrscheinlichkeit der Klassifizierung einer Nachricht als Hass verdoppeln kann [11]. Umgekehrt ergab eine finnische Studie, dass man nur

- Wörter falsch schreiben oder
- eine Zahl in einem W0rt einzuschmuggeln oder
- ganz einfach das Wort „Liebe“ irgendwo in einer Nachricht unterzubringen braucht,

um die Wahrscheinlichkeit ihrer Klassifikation als „Hass“ deutlich zu vermindern [3].

Williams und Mitarbeiter sehen diese Problematik durchaus, rechtfertigen ihre Arbeit zu Predictive Policing jedoch aufgrund der folgenden Vorteile: „Unsere Arbeit mildert die Gefahren vorhersagender Polizeiarbeit auf 3-fache weise ab:

- Die zur Abschätzung von Mustern verwendeten Daten werden nicht von der Polizei produziert und sind daher immun gegenüber den Vorurteilen, die dieser offizielle Generierungsprozess solcher Daten beinhaltet;
- Daten aus sozialen Online-Medien werden in Echtzeit gesammelt, wodurch Fehler vermieden werden, der durch ‚alte’ Daten entstehen, die den tatsächlichen Gegebenheiten nicht mehr entsprechen; und
- werden Minderheiten als mögliche Opfer und nicht als Täter betrachtet […]“[12] [19].

Der Versuch, Hass-Sprache zu erkennen und ihr entgegenzutreten, erscheint vor allem im gegenwärtigen Großbritannien notwendig, wie ein im New Scientist schon vor einem Jahr erschienener Kommentar mit dem Titel „Curbing hate speech isn’t censorship – it’s the law“ („Das Bändigen von Hass-Reden ist keine Zensur – sondern Gesetzt“) klar zum Ausdruck bringt: „Es mag eine interessante Debatte sein, ob Hass-Sprache – wie in der US-amerikanischen Verfassung – als Ausdruck von freier Rede geschützt werden sollte. Aber zum gegenwärtigen Zeitpunkt ist dies in Großbritannien nicht der Fall. Hass-Sprache ist vielmehr eine Straftat […] Das Ausspielen der Karte „Schutz der freien Rede“ wird oft zu einem Trojanischen Pferd, mit dem abscheuliche und reaktionäre Gedenken in unsere Gesellschaft eingeschmuggelt werden. Sofern [beispielsweise] die britischen Universitäten tatsächlich aktive Ideen-Zensur betreiben würden, haben Leute tatsächlich Grund zur Sorge, denen am Ausdruck des freien Willens gelegen ist. Aber die wahre Bedrohung der Werte der Aufklärung kommt von denen, die scheinheilig vorgeben, dass politische Korrektheit [und damit Zensur] zu weit gegangen sei. Die Verrücktheit liegt in der exakt gegenteiligen Richtung“[13] [1].

Es wäre in der Tat verrückt, nichts gegen Hass zu tun. Dass ein Rechtsstaat hier behutsam und mit Augenmaß vorgehen muss, um berechtigte Kritik die Freiheit der Rede nicht zu gefährden, ist offensichtlich. Erleichtert wird ihm dies durch unabhängige Wissenschaftler, die die Möglichkeit haben, Daten aus digitalen Medien zu analysieren und zu interpretieren, um Aufklärungsarbeit zu leisten. Von allen Bürgern muss man Besonnenheit fordern und vor allem

11 „The dashboard flags between 500 000 and 800 000 tweets per day related to Brexit, of which between 0.2 per cent and 0.5 per cent are classifed as hateful. About 0.2 percent of these are from users tagged with city locations within the UK, which the dashboard presents as a map of hate hotspots. If there is a spike, the information can be passed by analysts to the relevant local police forces.“

12 „Our work mitigates some of the existing pitfalls in prediction efforts in three ways: (1) The data used in estimating patterns are not produced by the police, meaning they are immune from inherent biases normally present in the official data generation process; (2) social media data are collected in real-time, reducing the error introduced by ‘old’ data that are no longer reflective of the context; and (3) viewing minority groups as likely victims and not offenders […]“ [19].

13 „Whether hate speech ought to be protected as free speech – as it is by the US constitution – is an interesting debate. But at present, in the UK, it is not. Hate speech is a criminal offence. […] Playing the free-speech card can often be a Trojan Horse for smuggling some deeply unpleasant and reactionary ideas back into society. If UK universities really were actively censoring ideas, then people who care about freedom of expression ought to be worried. But the true threat to enlightenment values comes from those who piously pretend that political correctness has gone too far. The madness lies in the exact opposite direction“ [1].

die Fähigkeit und das Bemühen, Hass von Kritik zu unterscheiden und die Freiheit zu verteidigen.

Literatur

[1] Anonymus. Curbing hate speech isn't censorship – it's the law. UK universities are being accused of suppressing ideas. All they are doing is complying with the law – and common decency. New Scientist 21.2.2918 www.newscientist.com/article/mg23731662-900-curbing-hate-speech-isnt-censorship-its-the-law/; abgerufen am 11.9.2019

[2] Bode F, Stoffel F, Keim D. Variabilität und Validität von Qualitätsmetriken im Bereich von Predictive Policing. Konstanzer Online-Publikations-System (KOPS) 2017 http://nbn-resolving.de/urn:nbn:de:bsz:352-0-402496; abgerufen am 11.9.2019

[3] Gröndahl T, Pajola L, Juuti M, et al. All You Need is "Love": Evading Hate Speech Detection arXiv:1808.09115v3 [cs.CL] 5 Nov 2018; abgerufen am 12.9.2019

[4] Hölig S, Hasebrink U. Reuters Institute Digital News Survey 2017: Ergebnisse für Deutschland. Arbeitspapiere des Hans-Bredow-Instituts Nr. 42. Hamburg: Verlag Hans-Bredow-Institut 2017

[5] Kaufmann M, Egbert S, Leese M. Predictive Policing and the Politics of Patterns. British Journal of Criminology 2019; 59: 674–692

[6] Knobloch T. Vor die Lage kommen: Predictive Policing in Deutschland. Chancen und Gefahren datenanalytischer Prognosetechnik und Empfehlungen für den Einsatz in der Polizeiarbeit. Stiftung Neue Verantwortung e. V., Berlin & Bertelsmann Stiftung, Gütersloh 2018 www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/predictive.policing.pdf; abgerufen am 11.9.2019

[7] Lu D. Google's hate speech-detecting AI appears to be racially biased. New Scientist 2019; 3243, 17.8.2019 (www.newscientist.com/article/2213064-googles-hate-speech-detecting-ai-appears-to-be-racially-biased/; abgerufen am 9.9.2019

[8] Lu D. UK police are using AI to spot spikes in Brexit-related hate crimes. New Scientist 2019; 3245, 31.8.2019 www.newscientist.com/article/mg24332453-500-uk-police-are-using-ai-to-spot-spikes-in-brexit-related-hate-crimes/; abgerufen am 9.9.2019

[9] Müller K, Schwarz C. Fanning the Flames of Hate: Social Media and Hate Crime (November 30, 2018). Social Science Research Network, SSRN https://ssrn.com/abstract=3082972 oder http://dx.doi.org/10.2139/ssrn.3082972; abgerufen am 9.9.2019

[10] Müller K, Schwarz C. From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment (July 3, 2019). Social Science Research Network, SSRN https://ssrn.com/abstract=3149103 oder http://dx.doi.org/10.2139/ssrn.3149103; abgerufen am 10.9.2019

[11] Sap M, Card D, Gabriel S, et al. The Risk of Racial Bias in Hate Speech Detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, S. 1668–1678. Florenz, Italien, 28.7–2.8. 2019 https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf; abgerufen am 11.9.2019

[12] Shearer E, Matsa KE, Pew Research Center. News Use Across Social Media Platforms 2018. Technical report file:///Users/manfredspitzer/Downloads/PJ_2018.09.10_social-media-news_FINAL.pdf; abgerufen am 9.9.2019

[13] Shetty S. Technology: force for progress, or tool of repression? Vortrag des [damaligen] Generalsekretärs von Amnesty International am Indian Institute of Technology (IIT) in Bombay on 16. Dezember 2016 https://www.amnesty.org/en/latest/news/2016/12/salil-shetty-speech-techfest/; abgerufen am 11.9.2019

[14] Spitzer M. Groß in Facebook, klein im Gehirn? Gehirnforschung zu sozialen Netzwerken. Nervenheilkunde 2012; 31: 299–304

[15] Spitzer M. www (WeltWeite Werbung) und die Folgen. Radikalisierung, Spionage, Vertrauens- und Wahrheitsverlust. Nervenheilkunde 2018; 37: 303–311

[16] Spitzer M. Smartphone und Depression: Ursache oder Therapie? Nervenheilkunde 2018; 37: 7–15

[17] Statista. Anzahl der Gemeinden in Deutschland nach Gemeindegrößenklassen, Stand 31.12.2017 https://de.statista.com/statistik/daten/studie/1254/umfrage/anzahl-der-gemeinden-in-deutschland-nach-gemeindegroessenklassen/; abgerufen am 10.9.2019

[18] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018; 359: 1146–1151

[19] Williams ML, Burnap P, Javed A, et al. Hate in the machine: Anti-Black and anti-muslim social media posts as predictors of off-line racially and religiously aggravated crime. Bristish journal of Criminology 2019, doi:10.1093/bjc/azz049 https://academic.oup.com/bjc/advance-article-abstract/doi/10.1093/bjc/azz049/5537169; abgerufen am 9.9.2019