

## **Briefing: Regulierungsalternativen im Umgang mit Hatespeech & Fakenews**

### **1) Ausgangspunkt**

Seit vergangenem Jahr steht das Löschen von als Hatespeech gemeldeten Inhalten nach Prüfung innerhalb von 24 Stunden durch Social-Media-Plattformanbieter (Facebook, Twitter, Google) in Form einer Selbstregulierung im Fokus des Vorgehens gegen Hatespeech im Netz. Den deutschen Kontext bietet die vom BMJV eingerichtete Taskforce<sup>1</sup> sowie das auf EU-Ebene eingerichtete Internet Forum, welches mit einem Code of Conduct<sup>2</sup> vergleichbare Maßnahmen ermöglicht. Seit Dezember 2016 steht auch ein Vorgehen gegen Fakenews durch das Bundesministerium des Innern (BMI) zur Debatte und wird in der Diskussion berücksichtigt.<sup>3</sup>

Netropolitische Kritik betraf bislang v.a. die Privatisierung der Rechtsdurchsetzung durch eine intransparente<sup>4</sup> Einschränkung der Meinungsfreiheit, die sich durch Unklarheiten der rechtlichen Eingrenzung potentiell illegaler Inhalte verstärkte<sup>5</sup>, sowie der antizipierten negativen Auswirkung auf eine potentielle Strafverfolgung<sup>6</sup>: Wie können gelöschte, strafbare Inhalte geahndet werden? Wie kann sicher gestellt werden, dass nur strafbare, nicht aber legale Inhalte voreilig gelöscht werden?

Mangelnde Kenntnisse über das tatsächliche Aufkommen von Hatespeech und Fakenews lassen die Wirkung bisheriger Maßnahmen schwer einschätzen, kennzeichnet die bisherige „Löschquote“ doch weder die korrekte Bearbeitung von Meldungen strafbarer Inhalte noch Erfolge ihrer Eindämmung.<sup>7</sup> Die Dynamik von Inhalten im Netz sowie der Wechsel einschlägiger Symbole und Plattformen impliziert eine geringe Wirksamkeit, der Hetze durch das Löschen von Inhalten beizukommen.

Gleichfalls unberücksichtigt scheinen bislang die unterschiedlichen Funktionsweisen der Plattformen und ihrer Nutzung, die alternative Regulierungsoptionen ermöglichen. Diese sind in der die US-amerikanischen Wahl folgenden Debatte um die Verantwortlichkeit sozialer Plattformen neuartig diskutiert worden und werden im Folgenden exemplarisch vorgestellt. Dem voran gestellt ist ein kurzer Abschnitt zur Verantwortlichkeit von Facebook im Kontext der Debatte, einem Sonderfall.

### **2) Hintergrund: Facebook und die Verbreitung unliebsamer Inhalte im Netz**

Die Bildung so genannter Echokammern im Netz ist breit diskutiert worden, allerdings zumeist in Bezug auf den Anteil der Nutzer an seiner Herstellung (Nachrichtenselektion, Likes, etc.).<sup>8</sup> Vernachlässigt wurde die Manipulation der Diskurse durch Facebook selbst, die über eine Verstärkung beliebter („relevanter“) Inhalte hinaus geht und auf Basis der systematischen Reduktion der Vielfältigkeit der den Nutzern angezeigten Inhalten die gesellschaftliche Polarisierung befördert.

Die Varietät dieser Inhalte ergibt sich durch a) den Newsfeed, b) die Plazierung der Inhalte im Newsfeed (zum Teil werbeeinflusst) und c) das Klickverhalten der Nutzer (durch Plazierung beeinflusst). Einer Facebook-Studie<sup>9</sup> zufolge reduziert der Newsfeed die Varianz der Inhalte bei liberalen Nutzern um 8%, bei konservativen Nutzern dagegen um 5%. Die Reduktion der Varianz durch das Klickverhalten läge bei liberalen Nutzer bei 6%, bei konservativen Nutzern bei 17%.

Die Beeinflussung des Klickverhaltens durch entsprechende Werbung (Plazierung) verbleibt in der Studie ebenso unklar wie die Beeinflussung der 96% Nutzer, die sich selbst nicht politisch identifizieren oder Facebook selten besuchen.<sup>10</sup> Allerdings zeigt die Studie, dass Facebook bereits unabhängig von konkreten Inhalten/ Werbung und Likes eine intransparente Verstärkung divergenter Diskursräume fördert, die Radikalisierung, Hatespeech und Fakenews den Boden bereitet.<sup>11</sup>

Darüber hinaus zeigte die US-Wahl, wie Facebook mittels der Erstellung von Psychogrammen, eigenen Daten und zusätzlich erworbenen persönlichen Daten eine sehr gezielte Wählermanipulation

auf Basis von Micro-Targeting durchführen konnte.<sup>12</sup> Informationen über zusätzlich zu den eigenen Daten erworbene Datenbanken bleiben unklar.<sup>13</sup>

Unter Berücksichtigung des US-Wahlkampfes stellt sich die Frage neu: Ist das Unternehmen, das auf vielfältige und intransparente Weise gesellschaftliche Diskurse und Wahlen beeinflusst, der richtige Adressat für die Entscheidung über die Einschränkung von Grundrechten (Meinungsfreiheit)? Bergen nicht die LösCHForderungen gegenüber Facebook, insbesondere mit Blick auf eine mögliche Haftung/ Bußgelder, die Gefahr der Verstärkung des Monopols, mit zweifelhafter Wirkung auf die Ziele: Die Verfolgung rassistischer und anderer strafbarer Inhalte zum Schutz der Demokratie?

### 3. Sozio-technische Alternativen im Umgang mit Hatespeech und Fakenews

Die allgemeine Manipulation des Diskurses, die Polarisierung der Wählerschaft insbesondere durch Facebook, macht eine **Transparenz über Algorithmen** notwendig, insbesondere mit Blick auf die Newsfeeds/ Timelines (aktueller US-Diskurs<sup>14</sup>). Darüber hinaus: Wäre ein Verbot der Beeinflussung der personalisierten Nachrichten durch Werbung denkbar, gehen doch viele Nutzer von einer Objektivität „ihrer“ Nachrichten aus? Oder die Verpflichtung auf anwendungsfreundliche Transparenzoptionen, z.B. Auswahl zwischen einem durch Werbung beeinflussten und unbeeinflussten Newsfeed? Bieten sich hier Anreize für Unternehmen (Reduzierung von Steuern, Befreiung von Haftbarkeit)?<sup>15</sup>

Die Gefahr der Manipulation von Wahlen durch (vom Newsfeed separates) Micro-Targeting rückt den **Datenschutz** (das Zweckbindungsgebot, das Verbot einer entsprechenden Profilbildung) ins Zentrum: Mittelfristig stehen beide Elemente im Zuge der Umsetzung der EU-Datenschutz-Grundverordnung zur Debatte, der Entwurf des Datenschutz-Anpassungs- und Umsetzungsgesetzes (DSAnpUG-EU) des BMI schwächt diese leider erheblich<sup>16</sup> - bedauerlich, verweist doch hier gleichfalls die amerikanische Debatte auf die Problematik der Zweckentfremdung von Daten.<sup>17</sup>

Es folgen konkrete Maßnahmen gegen Hatespeech und Fakenews:

#### a) Das Monitoring: Zielgerichtete Intervention, Strafverfolgung & Evaluation

Mit Gefahr des „Überspringens“ von Hatespeech in reale Gewalt im Kontext nationaler Wahlen konfrontiert, entwickelte das Project iHub in Kenya offene, mit Automatisierung und maschinellem Lernen kombinierte Monitoring-Technologien.<sup>18</sup> Ziel war die rechtzeitige Intervention bei Gewalt. Die Technologien werden aktuell in Nachbarstaaten getestet, Evaluationen stehen im Januar 2017 an.

Das Monitoring kann zunächst Aufschluss geben zum Aufkommen von Hatespeech und Fakenews – und ob Maßnahmen zu seiner Reduktion beitragen. Einsichten in die **Dynamik** der Verbreitung von Hatespeech/ Fake News – über Verursacher, Adressaten, Knotenpunkte etc. - ermöglichten in Kooperation mit Strafverfolgungsbehörden eine **strategische Ahndung**: Durch die Berücksichtigung des Einflusses der Speaker könnte bspw. die Überforderung der Behörden in der Verfolgung strafbarer Inhalte begrenzt, die abschreckende Wirkung entsprechender Urteile maximiert werden. Eine Kooperation mit Polizei und Sicherheitskräften erlaubte eine **rechtzeitige Intervention bei Gewalt**.

In Verbindung mit einer Prüfung von Hatespeech-Meldungen könnte das Monitoring darüber hinaus den **Umgang mit LösCHForderungen** gegenüber den Plattformen transparenter gestalten, der Gefahr des übermäßigen LösChens von Inhalten durch voraus eilenden Gehorsam (z.B. in Folge möglicher Haftung/ Bußgelder) begegnen.

Allerdings ist die **Transparenz und Unabhängigkeit** des Monitorings von entscheidender Bedeutung. Sie unterscheidet die Maßnahme bspw. von den prominent kritisierten Plänen der Massendatenauswertung von Internetinhalten durch das Bundesamt für Verfassungsschutz (BfV).<sup>19</sup>

### **Exkurs: Zur Bekämpfung der Verbreitung von Fakenews auf Facebook in den USA**

Einer kurzen Diskussion verschiedenster Optionen<sup>20</sup> infolge der US-Wahl folgte ein Maßnahmenbündel seitens Facebook, bei dem Nutzer zunächst Fakenews melden können, die daraufhin von einer unabhängigen Institution geprüft werden. Sollte sich der Verdacht einer Falschmeldung bewahrheiten, werde die Meldung entsprechend markiert (mit Links zu Erläuterungen). Seine Weiterverbreitung werde durch Rückfragen und ein Downgrading im Newsfeed behindert. Zudem sollen finanzielle Anreize zur Verbreitung von Fake News eingedämmt werden. Die Prüfungsinstitutionen arbeiten finanziell unabhängig.<sup>21</sup>

Einerseits mangelt es der Maßnahme an Transparenz im Bereich der eingeschränkten Verbreitungsmöglichkeiten (Newsfeed, Werbung), die nicht zuletzt Vergleiche mit China hervorruft.<sup>22</sup> Andererseits stellt die externe Prüfung markierter Inhalte einen Fortschritt hinsichtlich der Transparenz von Eingriffen in die Meinungsfreiheit (Verfahren) dar, betont die **Bedeutung unabhängiger, transparenter Checks & Balances** und markiert einen eindeutigen Wandel in der amerikanischen Regulierungsdebatte.

### **b) Wo Twitter aktuell voran geht: Das technische Design konstruktiver Diskurse**

Neben der Bekämpfung der Verbreitung von Hatespeech und Falschmeldungen im Netz stellt sich die Frage: Wie finden wir zurück zu einem konstruktiven Diskurs? Oder: Wie kommen wir zu konstruktiven Ergebnissen, wenn alle am digitalen Diskurs teilhaben dürfen?

Hier bieten neuere, technische Design-Ansätze wertvolle Beiträge. Sie eint zunächst die Definition einer positiven Vision von Community (Welchem Zweck soll der Austausch auf der Plattform dienen, warum sollen Menschen miteinander kommunizieren?)<sup>23</sup> sowie klarer Regeln des Diskurses. Zugrunde liegen Studien der Bedürfnisse und Dynamiken der unterschiedlichen Online-Communities wie sie bspw. das Coral Project in New York, eine Kooperation zwischen Mozilla, der New York Times und der Washington Post, oder Caroline Sindors (BuzzFeed/ IBM Watson) durchführten.<sup>24</sup>

Daraus abgeleitet ergibt sich bspw. die Möglichkeit einer fundierte Argumente und Speaker unterstützenden, transparenten Kommentar-/ Diskursmoderation.<sup>25</sup> Oder etwa die Gestaltung individueller Diskursräume durch eine Ausdifferenzierung der Privatheitseinstellungen im Kontext konkreter Inhalte. Hierzu zählen etwa die das Muten („Stummschalten“) einzelner Hashtags oder Gespräche, wie Twitter sie jüngst integrierte.<sup>26</sup> Sie erlauben dem Nutzer, sich selbst vor virulenten Diskussionen zu schützen ohne die Plattform verlassen zu müssen. Anwendungsbeispiele für Facebook zielen bspw. auf Privatheitseinstellungen der Kommentare, die einzelne Nutzer schützt und Bewusstsein über die Öffentlichkeit von Kommentaren schafft (technisch unterstützte Medienkompetenz). Da die Lösungsansätze stark nach den Plattformen variieren, müssten Expertengespräche geführt werden.

Die neue Plattform Imzy aus dem Twitter/Reddit-Umfeld integriert bspw. verschiedene Ansätze: Sie verbindet umfangreiche Editierfunktionen, eine anwendungsfreundliche Kommentarmoderation sowie ausdifferenzierte Privatheitseinstellungen miteinander. „It's like Reddit, Without the Trolls“, so Bloomberg zur neuen Facebook-Alternative, die der Notwendigkeit eines konstruktiven Diskurses und den Funktionen sozialer Netzwerke (Ausbreitung und Virulenz) gerecht wird.<sup>27</sup>

Inwiefern diese Design-Ansätze zum Tragen kommen können, hängt einerseits ab von regulativen Spielräumen (Haftung/ Bußgelder schaden Facebook-Konkurrenten<sup>28</sup>) sowie der Einbindung durch große Plattformen, die wie bei Transparenzpflichten auch durch Anreize erhöht werden könnte. Völlig unklar bleibt an dieser Stelle, warum Twitter seinen positiven Zugang zu Design-Optionen im Kampf gegen Hatespeech und Belästigung nicht entsprechend kommuniziert – sondern den Eindruck verstärkt, Hatespeech sei egal (geringe Löschquote). Die Entwicklung der Design-Ansätze zeigen dagegen an, dass Hatespeech auch für Plattformen wie Twitter ein Thema ist.

- 1 BMJV (15.12.2016): „Gemeinsam gegen Hassbotschaften – Task Force stellt Ergebnisse vor“, [https://www.bmjbv.de/SharedDocs/Artikel/DE/2015/12152015\\_ErgebnisrundeTaskForce.html](https://www.bmjbv.de/SharedDocs/Artikel/DE/2015/12152015_ErgebnisrundeTaskForce.html).
- 2 European Commission (31.05.2016): „European Commission and IT Companies announce Code of Conduct on illegal online hate speech“, [http://europa.eu/rapid/press-release\\_IP-16-1937\\_en.htm](http://europa.eu/rapid/press-release_IP-16-1937_en.htm).
- 3 Knaup, Horand/ Traufetter, Gerald (23.12.2016): „Innenministerium will Abwehrzentrum gegen Falschmeldungen einrichten“, <http://www.spiegel.de/netzwelt/netzpolitik/fake-news-bundesinnenministerium-will-abwehrzentrum-einrichten-a-1127174.html>.
- 4 Krause, Till/ Grassegger, Hannes (16.12.2016): Facebook’s secret rules of deletion: <http://international.sueddeutsche.de/post/154543271930/facebooks-secret-rules-of-deletion>.
- 5 Fleischhauer, Jan (03.10.2016): Die Hassmaschine: <http://www.spiegel.de/politik/deutschland/fleischhauer-kolumne-hassmaschine-facebook-a-1114985.html>.
- 6 Beckedahl, Markus (26.10.2016): „Facebook löscht 100.000 Inhalte aus Deutschland, Bundesregierung weiß von nichts“, <https://netzpolitik.org/2016/facebook-loescht-100-000-inhalte-aus-deutschland-bundesregierung-weiss-von-nichts/>.
- 7 Maas, Heiko (09.01.2017) im Interview: „Die Meinungsfreiheit hat auch ihre Grenzen“, <http://www.medienpolitik.net/2017/01/medienpolitik-die-meinungsfreiheit-hat-auch-grenzen/>.
- 8 Zusammenfassend beispielsweise: Stöcke, Christian (25.01.2016): „Radikal dank Facebook“, <http://www.spiegel.de/netzwelt/netzpolitik/filterblase-radikalisierung-auf-facebook-a-1073450.html>.
- 9 Bakshy, Eytan et al. (07.05.2015): „Exposure to ideologically diverse news and opinion on Facebook“, <http://science.sciencemag.org/content/early/2015/05/06/science.aaa1160>; Sandvig, Christian (07.05.2015, updated): „The Facebook „It’s Not Our Fault“ Studie: <https://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/>“.
- 10 Tufekci, Zeynep (07.05.2015): „How Facebook’s Algorithm Suppresses Content Diversity (Modestly) and How the Newsfeed Rules Your Clicks“, <https://medium.com/message/how-facebook-s-algorithm-suppresses-content-diversity-modestly-how-the-newsfeed-rules-the-clicks-b5f8a4bb7bab#.u4qef46uz>; Jurgenson, Nathan (07.05.2015): „Facebook: Fair and Balanced“, <https://thesocietypages.org/cyborgology/2015/05/07/facebook-fair-and-balanced/>.
- 11 Baer, Drake (09.11.2016): „The ‚Filter Bubble‘ Explains Why Trump Won and You Didn’t See It Coming“, <http://nymag.com/scienceofus/2016/11/how-facebook-and-the-filter-bubble-pushed-trump-to-victory.html>.
- 12 Beispielsweise seien ängstlichen Menschen mit hohen Neurotizismus-Werten (emotionale Labilität/ Ängstlichkeit) Waffen als Versicherung verkauft worden, unterstützt durch Bilder von Einbrüchen. Konservativen Menschen mit hoher Extraversion (Geselligkeit) wurden dagegen positive Bildern von Jagdsituationen im Sonnenuntergang gezeigt. Hintergrund: Krogerus, Mikael (03.12.2016): „Ich habe nur gezeigt, dass es die Bombe gibt“, <https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-zeigt-dass-es-die-bombe-gibt/>.
- 13 Angwin, Julia et al. (27.12.2016): „Facebook Doesn’t Tell Users Everything It Really Knows About Them“, <https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them>.
- 14 Coren, Michael J. (06.12.2016): „Facebook needs to hand over its algorithm if it really wants to end fake news“: <https://qz.com/847640/facebook-needs-to-hand-over-its-algorithm-if-it-really-wants-to-end-fake-news/>; Garside, Juliette (18.12.2016): „Labour calls for closer scrutiny of tech firms and their algorithms“, <https://www.theguardian.com/business/2016/dec/18/labour-calls-for-regulation-of-algorithms-used-by-tech-firms>; siehe auch Ressourcensammlung [https://www.imzy.com/smart\\_citizens/post/value\\_faults\\_of\\_artificial\\_intelligence](https://www.imzy.com/smart_citizens/post/value_faults_of_artificial_intelligence).
- 15 Zittrain, Jonathan (2014): „Engineering an Election“: [https://papers.ssrn.com/sol3/Papers.cfm?abstract\\_id=2457502](https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2457502); Zusammenfassung: Zittrain, Jonathan (02.06.2014): „Facebook could decide an Election Without Anyone Ever Finding Out“, <https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering>.
- 16 Schaar, Peter (03.12.2016) im Interview: „Neues Datenschutzgesetz soll uferlose Datenauswertung ermöglichen“, <https://netzpolitik.org/2016/interview-mit-peter-schaar-neues-datenschutzgesetz-soll-uferlose-datenauswertung-ermoeglichen/>.
- 17 Zittrain, Jonathan (2014): „Engineering an Election“: [https://papers.ssrn.com/sol3/Papers.cfm?abstract\\_id=2457502](https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2457502); Zusammenfassung: Zittrain, Jonathan (02.06.2014): „Facebook could decide an Election Without Anyone Ever Finding Out“,



<https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering>.

- 18 Sambuli, Nanjira (10.03.2016): „The importance of monitoring online hate speech“, <http://www.dw.com/en/the-importance-of-monitoring-online-hate-speech/a-19104789>; Mit Informationen zur Entwicklung des Projektes: Sambuli, Nanjira (19.08.2014): „Umati’s Tech Evolution“, <http://ihub.co.ke/blogs/20152> (neue Berichte ab Januar 2017 zu erwarten).
- 19 Meister, Andre (15.04.2015): „Geheime Referatsgruppe: Wir enthüllen die neue Verfassungsschutz-Einheit zum Ausbau der Internetüberwachung (Updates)“, <https://netzpolitik.org/2015/geheime-referatsgruppe-wir-praesentieren-die-neue-verfassungsschutz-einheit-zum-ausbau-der-internet-ueberwachung/>.
- 20 Beispielsweise: Woolf, Nickey (29.11.2016): „How to solve Facebook’s fake news problem: experts pitch their ideas“, <https://www.theguardian.com/technology/2016/nov/29/facebook-fake-news-problem-experts-pitch-ideas-algorithms>; Scola, Nancy (12.11.2016): „Fact checkers to Facebook: Fight fake news“, <http://www.politico.com/story/2016/12/facebook-fake-news-fact-checkers-232444>.
- 21 Jamieson, Amber/ Solon, Olivia (15.12.2016): „Facebook to begin flagging fake news in response to mounting criticism“, <https://www.theguardian.com/technology/2016/dec/15/facebook-flag-fake-news-fact-check>; in der deutschen Zusammenfassung Reuter, Markus (15.12.2016): „Facebook ergreift Maßnahmen gegen Fake News – in den USA“, <https://netzpolitik.org/2016/facebook-ergreift-massnahmen-gegen-fake-news-in-den-usa/>.
- 22 Custer, C. (20.12.2016): „In China, Facebook’s fake news crusade means something completely different“, <https://www.techinasia.com/china-facebooks-fake-news-crusade-means-completely>.
- 23 Community Visionen/ Reflektionen bspw. Kathrine Maher (29.10.2016): „Privacy and Harassment on the Internet“, [https://en.m.wikisource.org/wiki/Privacy\\_and\\_Harassment\\_on\\_the\\_Internet](https://en.m.wikisource.org/wiki/Privacy_and_Harassment_on_the_Internet).
- 24 Coral Project (2016): Research, <https://coralproject.net/research.html>; Caroline Sindere, Webseite: <http://carolinesinders.com/about/>; Übersicht: „RP17: Speaker Caroline Sindere tames the Internet“, <https://re-publica.com/en/17/news/rp17-speaker-caroline-sinders-tames-internet>; Talks e.g. (2016): „Fighting online harassment with good design“, <http://www.cbc.ca/radio/spark/320-wake-up-the-brain-with-sound-and-more-1.3568478/fighting-online-harassment-with-good-design-1.3573323>; „Online Harassment“, <https://videos.theconference.se/caroline-sinders-online-harassment>. Ebenfalls äußerst spannend: Welche Social Media Daten geben Auskunft worüber? Warum die lautesten Hater mitnichten die Mehrheit repräsentieren? Dyer, Harry T. (09.12.2016): „What social media data should I use in my research? A response to Choi et al. (2016)“, <https://harrydyer.com/2016/12/09/what-social-media-data-should-i-use-in-my-research/>.
- 25 Coral Project (2016): Products, <https://coralproject.net/#products>.
- 26 Caroline Sindere (2016): „Fighting online harassment with good design“, <http://www.cbc.ca/radio/spark/320-wake-up-the-brain-with-sound-and-more-1.3568478/fighting-online-harassment-with-good-design-1.3573323>; Implementation bei Twitter: Klonick, Kate (15.11.2016): „Twitter Releases New Features to Fight Harassment – Just in Time“, [http://www.slate.com/blogs/future\\_tense/2016/11/15/twitter\\_releases\\_new\\_features\\_to\\_fight\\_harassment.html](http://www.slate.com/blogs/future_tense/2016/11/15/twitter_releases_new_features_to_fight_harassment.html).
- 27 Stein, Joel (07.12.2016): „It’s like Reddit, Without the Trolls“, <https://www.bloomberg.com/news/articles/2016-12-07/it-s-like-reddit-without-the-trolls>.
- 28 Die Frage, die sich hier stellt und welche zu überprüfen ist: Wer kann sich die kostenintensive Prüfung von Inhalten leisten? Ähnlich der Urheberrechtsdebatte (Uploadfilter) bergen Haftung und Bußgelder die Gefahr, bessere Konkurrenten in ihrer Ausgangsposition zu benachteiligen und damit die Verstärker von Hatespeech und Fakenews (v.a. Facebook) indirekt zu stärken.

Alle Quellen wurden zwischen am 11.01.17 überprüft.