Trolle 2.0
Eine exemplarische Analyse von Hate Speech in den sozialen Netzwerken

Viola Melzner

Abstract: Im öffentlichen Kommunikationsraum sozialer Netzwerke werden komplexe soziale Konflikte auf das eindeutig lesbare Oppositionspaar *Like* / *Dislike* reduziert. Als relevant werden hierbei jene Inhalte gewertet, die viele Reaktionen generieren – positive wie auch negative. Diese Umgebung bietet eine große Bühne für Trolle, die sich mittels provokanter und pejorativer Sprache Gehör verschaffen. Die Zuschreibungen an diese Störenfriede sind veränderlich. Ein Bild aggressiver, psychopatischer Einzeltäter ist veraltet. In einer Zeit, in der sich immer mehr User*innen in verschiedenen Kontexten in einer Art Ansteckungseffekt dem vorherrschenden Duktus anschließen, braucht es einen neuen Trollbegriff. Trolle 2.0 sind User*innen, die sich – je nach Gemütslage und kurzfristigem Ziel, bewusst oder unbewusst – dem diffamierenden Umgangston in sozialen Netzwerken anschließen. Hierbei müssen auch einmalige emotionale Entladungen eingeschlossen werden. Nur so kann Hate Speech in sozialen Netzwerken als gesamtgesellschaftliches Phänomen erfasst werden. Denn die unheilvolle Wirkung der massenhaften Verwendung von Hate Speech durch Trolle 2.0 liegt darin, dass Hass zum gängigen Umgangston lanciert.

Zur Person: Viola Melzner studierte MA Allgemeine und Vergleichende Medienwissenschaft an der Universität Regensburg. Der vorliegende Beitrag basiert auf ihrer Masterarbeit. Betreuerin: Prof. Dr. Christiane Heibach.

Schlagwörter: Trolle; Hate Speech; soziale Netzwerke; Facebook; Web 2.0

Das Web 2.0 verbindet Nutzer*innen weltweit. Dabei motiviert besonders der Teilbereich des Social Web dazu, Informationen, Gedanken und Gefühle (mit) zu teilen (John, 2017: 52). Es entsteht auf diese Weise eine Pluralität an Stimmen, der enthusiastische Einschätzungen ein hohes demokratisierendes

Potenzial beimessen, denn eine hermetisch abgeriegelte, staatlich kontrollierte Infosphäre werde so verhindert (Wilson et. al., 2013). Doch die Hoffnungen scheinen gegenwärtig ins Gegenteil umzuschlagen, da sich politische Debatten online verschärfen: Social Bots beeinflussen das Meinungsklima; der sogenannte Islamische Staat kommuniziert seine propagandistischen Narrative (Ebner, 2018: 154 f.); rechtspopulistische und rechtsextreme Gruppen vernetzen sich (Schaeffer, 2018: 25); Echokammern verdichten sich, und Populisten säen Dissens und propagieren einfache Lösungen für komplexe Probleme (Ebner, 2018: 166).

Auch Umgangsformen und Gesprächskonventionen sind betroffen. So wird Hate Speech im Internet zur perfiden Alltags- oder auch Allzweckwaffe (Fleischhack, 2017: 27), um gegensätzliche Meinungen zu eliminieren. Während Hate Pages und Hassforen anfangs Nischenerscheinungen des Internets waren, werden sie in sozialen Netzwerken zum Alltag. Julia Ebner (2018: 31) fasst diesen Zustand mit dem Begriff „Zeitalter der Wut" zusammen.

Die zugrundeliegende Kommunikationsform bildet ein schwer fassbares Feld. Manche Fälle von Hate Speech lassen sich schnell – teils auch vorschnell – identifizieren. Besonders wenn Ethnophaulismen oder Schimpfworte verwendet werden. In anderen Situationen fällt die Bewertung schwerer, wenn die Bedeutung von Begriffen unklar ist. Der ursprünglich abwertende Ausdruck *queer* beispielsweise erhielt eine Neubewertung und wurde zum Eigenverständnis der bezeichneten Gruppe. Judith Butler (2006: 29 f.) spricht dabei von diskursiver Performativität. In wieder anderen Fällen wird Hate Speech vorschnell angenommen, obwohl es sich schlicht um Kritik handelt.

Hate Speech wirft durch ihr massenhaftes Auftreten in sozialen Netzwerken nicht nur die Fragen auf, ob verbalisierter Hass durch Algorithmen befördert wird oder ob in den räumlich und zeitlich versetzten Dialogen die Hemmschwelle sinkt. Ebenso interessant ist, wer sich hinter den hasserfüllten Posts versteckt. Schnell werden in diesem Zusammenhang Begriffe wie *Hater* oder *Trolle* in den Ring geworfen, wobei Letztere im Folgenden genauer in den Blick genommen werden.

Tatsächlich bewohnen Trolle das World Wide Web seit dessen Aufkommen und stellen kein neues Phänomen dar. Doch ist gegenwärtig eine Radikalisierung des Trollwesens zu bemerken, die vor allem von der Verwendung von Hate Speech bestimmt ist und sich deutlich von spitzbübischen Streichen der Forentrolle in den 1990er Jahren unterscheidet. Aufgrund dessen kann von einer Weiterentwicklung des Trolls im Web 2.0 gesprochen werden, welche die Einführung eines neuen Begriffs notwendig macht: Die gegenwärtige Entwicklungsstufe soll im Folgenden als „Troll 2.0" bezeichnet werden.

Trolle 2.0 – Ihre Geschichte, ihr Habitat, ihr Wesen

1990 wurde im Rahmen einer Newsgroup, einem Vorläufer sozialer Netzwerke, erstmalig ein Internetstörenfried als Troll bezeichnet (Ley, 2018: 18). Der Begriff des *trollings* leitet sich, wie Ingrid Brodnig (2013: 92 f.) darlegt, aus der Anglersprache ab. Genauer: es handelt sich um die englische Bezeichnung des Schleppenfischens. Ebenso wie der Fischer werfe der Internettroll Köder aus, um andere Nutzer an die Angel zu bekommen.

Gleichzeitig bietet eine Benennung als Troll assoziative Bezugspunkte zu den gleichnamigen norwegischen und schwedischen Sagenwesen. In den skandinavischen Erzählungen, die ab Ende des 9. Jahrhunderts aufkamen, werden Trolle als todbringende, übernatürliche und gefährliche Wesen charakterisiert (Simek, 2018: 11 f.). Über die Zeit wurden die grausamen Wesen von Kinderbuchautoren zwar verharmlost und verniedlicht, doch im digitalen Kontext lebt die ursprüngliche Bedrohlichkeit der Trolle wieder auf. So gelten Internettrolle als reale Gefahr mit Konsequenzen für das menschliche Miteinander. Öffentliche Diskussionen werden von ihnen gestört, Gesprächskonventionen niedergerissen.

Zwar erlangte das Trollphänomen erst im Web 2.0 durch das sogenannte „RIP-Trolling“[1] öffentliche Aufmerksamkeit (Karppi, 2013: 280), doch bewohnen Trolle das Internet schon seit dessen Anfängen. Für den Forentroll der 90er Jahre stand vor allem das Amüsement der eigenen Gruppe im Vordergrund (Bishop, 2014: 9), indem sie technisch weniger erfahrene Nutzer*innen vorführten (Fichman / Sanfilippo, 2016: 44). Dass diese Umgangsform nicht zu gesellschaftlicher Besorgnis führte, sieht Ley (2018: 16) in einem scheinbar akzeptierten Klischee der Zeit begründet: „Es waren die Nerds, so das damals gängige Vorurteil, die Computerfreaks, die zu lange vor dem Bildschirm sitzen […] und dabei vergessen haben, wie man vernünftig miteinander umgeht.“ Doch während der vermeintliche Computernerd im Web 1.0 noch als Person hinter dem Forentroll identifiziert werden konnte, gestaltet sich eine eindeutige Zuordnung mittlerweile deutlich schwieriger. Das Internet ist praktisch jedem zu jeder Zeit zugänglich. Damit diversifiziert sich das Trollphänomen im Web 2.0 und die definitorischen Grenzen verschwimmen. Gleichzeitig kommt es zu einer Radikalisierung des Trollwesens: „The term trolling has essentially gone from meaning provoking others for mutual enjoyment to meaning abusing others for only one's own enjoyment.“ (Bishop, 2014: 8) Auch Ley (2018: 15) sieht seit der Jahrtausendwende eine Entwicklung hin zu einer Sprache voller Beschimpfungen und mangelnder Toleranz.

Diese inflationäre Ausbreitung einer radikalisierten Trollmanier muss definitorisch unter einen Begriff gefasst werden. Nur so kann das Phänomen angemessen beschrieben werden, und nur so können auch jene Nutzer*innen erfasst werden, die sich einmalig dem rauen Umgangston anschließen. Im

1 Beim RIP-Trolling werden die Facebookseiten von Verstorbenen spöttisch und höhnisch kommentiert (Karppi, 2013: 280).

Sinne einer evolutionären Entwicklung kann, in Anspielung auf das Habitat (das Web 2.0) die gegenwärtige Ausprägung als Troll 2.0 bezeichnet werden. Es geht bei dem Begriff vornehmlich um den Gesamteffekt, der erzielt wird, wenn eine Vielzahl an Nutzer*innen einen hasserfüllten Sprachgebrauch wählen. Daher muss sich die Untersuchung der Trolle 2.0 deutlich von frühen Forschungen abheben, welche die Motive der Trolle auf psychologische Merkmale der Täter zurückführen – das Onlineverhalten als Reaktion auf das Offlineleben; der grobe Umgang als Folge selbst erfahrener Schikanen. Denn im Lichte dieser Argumentation erscheint das Verhalten in der individuellen Persönlichkeit verankert, nicht durch Umwelt oder andere Einflüsse bedingt. Justin Cheng et. al. (2017: 2) postulieren hierzu: „That is, trolls are born, not made." Aus wissenschaftlicher Sicht ist diese Annahme zu problematisieren. Trolle 2.0 sind, so lässt der Umfang hasserfüllter Äußerungen vermuten, keine von der Norm abweichenden Einzeltäter. Eine derartige Argumentation schließt aus, dass der Nachbar von Nebenan dem Trolltypus entsprechen könnte, wenn dieser nicht das prototypische Bild eines bösartigen Psychopathen erfüllt. Folglich droht eine potentielle Verharmlosung einer sich ausbreitenden Umgangsform, ein blinder Fleck gegenüber einem aufkommenden Duktus des Hasses.

Allgemein lässt sich definieren: Ein Troll 2.0 ist jeder, der sich – je nach Gemütslage, je nach kurzfristigem Ziel, bewusst oder unbewusst – dem diffamierenden Umgangston in sozialen Netzwerken anschließt. Die durch verschwimmende Definitionsgrenzen erschwerte Identifikation von Trollen 2.0 erfolgt anhand folgender sieben Kriterien. Nicht alle beschreiben zwangsläufig die intendierten Ziele der Trolle 2.0, dennoch aber die Konsequenzen ihres Handelns:

(1) *Der Moment des Öffentlichen*: Im Kontext sozialer Netzwerke agieren Trolle 2.0 in der Öffentlichkeit. Ihre Äußerungen haben ein potenziell weltumfassendes Publikum. Die Resonanz anderer Nutzer*innen, ihre Verärgerung, ihre Empörung bildet die entscheidende Triebfeder für ihr Verhalten.

(2) *Gefühlte Anonymität und Distanz*: Der Troll 2.0 agiert im Glauben keine Konsequenzen für sein Handeln tragen zu müssen, keine Gesetze achten zu müssen und „die simpelsten Regeln des Miteinanders ignorieren zu dürfen" (Ley, 2018: 20).

(3) *Das eigene Amüsement*: Trolle 2.0 empfinden Freude an Uneinigkeiten und der Empörung anderer Gesprächsteilnehmer*innen (Fichman / Sanfilippo, 2016: 69). Oftmals wird dabei das rhetorische Mittel der Ironie als Vorwand und Rechtfertigung des eigenen Verhaltens genutzt. Der Slogan „for the lulz"[2] kann in diesem Sinne als eine Art Schlachtruf der Trolle 2.0 gewertet werden.

(4) *Das Spiel mit den Emotionen*: Trolle 2.0 handeln mit dem Ziel, die Gefühle anderer ins Chaos zu stürzen. Ley (2018: 20) spricht von einem „asoziale[n] Wesen in sozialen Netzwerken". Auf ihre Provokation hin werden

2 *lulz* ist laut Phillips Whitney eine Ableitung des Ausdrucks *lol* (*laughing out loud*) und beschreibt das Amüsement auf Kosten anderer (Milner, 2013: 66).

Verwirrung, Wut und Trauer ausgelöst (Brodnig, 2013: 93). In Anlehnung an den Journalisten Adrian Chen schreibt Brodnig (ebd.: 94): „Trolle sind quasi Hacker unserer Gefühle, sie versuchen, die Mechanismen des menschlichen Miteinanders zu durchsuchen, und greifen dort an, wo sie Schwachstellen finden."

(5) *Das Moment der Wiederholung*: Mit ihrem Verhalten versuchen Trolle 2.0 die Zeit und Energie anderer zu verschwenden (Fichman / Sanfillipo, 2016: 10). Probates Mittel hierzu ist die Wiederholung (Simek, 2018: 218). Sie beharren stur auf ihren Darstellungen und lassen Gegenargumente nicht gelten, gehen auf diese nicht weiter ein. Fichman und Sanfilippo (2016: 13) schreiben in diesem Zusammenhang auch, dass wiederkehrende *Hate Speech* aufgrund der Wiederholung zu *trolling* werde.

(6) *Die Dekonstruktion der Gepflogenheiten*: Trolle 2.0 verwehren sich den gängigen Regeln gemeinsamer Diskussionen. Sie verstehen es, Reizthemen zu entfachen, um die Grundsätze zivilisierter Kommunikation zu verletzen (Simek, 2018: 218). Laut Alexander Glück (2013: 12) stelle es für Trolle eine Auszeichnung dar, wenn sich Internetcommunities, die sich für Meinungsfreiheit einsetzen, dem Störenfried gegenüber zum Musterbeispiel für Intoleranz entwickeln.

(7) *Sprachliche Grobheit als Form des Ausdrucks*: Trolle 2.0 nutzen Hate Speech, um andere Gesprächsteilnehmer*innen zu diffamieren, einzuschüchtern und zu dehumanisieren (Whillock / Slayden, 1995: xiii).

Hate Speech – die Sprache der Trolle 2.0

Der Terminus Hate Speech verweist auf die Emotion Hass – ein Umstand der in der Literatur als irreführend kritisiert wird (z.B. Sponholz, 2018; Unger, 2013), denn eine hasserfüllte Gemütslage ist nicht zwingend notwendig, um Hate Speech zu kommunizieren. Vorstellbar sind beispielsweise rassistische Diskurse, in denen sich dehumanisierende Redewendungen konventionalisiert haben (Meibauer, 2013: 3). Um trügerischen Assoziationen entgegenzuwirken, entstanden im Medien- und Fachdiskurs daher zahlreiche nuancierte Bezeichnungen, wie beispielsweise „Netzhass" (Fleischhack, 2017), „Online-Hass" (ebd.), „extreme speech" (Pohjonen / Udupa, 2017) oder „Words That Wound" (Delgado / Stefancic, 2004).

Im Folgenden soll, um die Begriffsinflation nicht weiter zu treiben, kein neuer Terminus eingeführt werden. Zur genaueren Konturierung der Trolle 2.0 soll stattdessen ein Abgleich unterschiedlicher Definitionen erfolgen. Die so entstehende Synthese soll dazu dienen, diese Kommunikationsform als *modus operandi* von Trollen 2.0 zu präzisieren. Im Alltagsverständnis ist die Beurteilung eines Hate Speech-Vorfalls oftmals entscheidend von dessen Konsequenzen geprägt. Wird Hate Speech jedoch in den Kontext offener Gewalt gestellt, erlangt gewaltvolle Sprache – so scheint es – ausschließlich

dann eine verletzende Wirkung, wenn sie bestimmte Effekte hervorruft (Butler, 2006: 68).

Eine derartige Charakterisierung gilt es in dreierlei Hinsicht zu problematisieren. Erstens kann Hate Speech unter dieser Prämisse nur in Beurteilung konkreter Fälle eindeutig identifiziert werden. Hierzu müsste die Wirkung auf Betroffene festgestellt werden (Sirsch, 2013: 168).

Zweitens würde Hate Speech dann *per definitionem* alle Äußerungen umfassen, von denen sich jemand belästigt fühlt. Dann könnte laut Unger (2013: 274) „auch der *Christopher Street Day* mit dem Verweis auf die homophobe Nachbarschaft, die sich von der Parade stark belästigt fühlt und ihr nicht ausweichen kann, verboten werden."

Drittens erscheint physische Gewalt in dieser Argumentation als Kausalfolge von Hate Speech. Die verbale Herabwürdigung stellt jedoch keine Determinante für Gewaltausschreitungen dar. Dennoch wohnt einer Einschüchterung beispielsweise eine verletzende Wirkung inne, auch dann, wenn die angedrohten Folgen nicht eintreten. Denn die von einer Drohung ausgehende Gefahr besteht darin, dass sie verbal ankündigt, was der Körper tun *könnte* (Butler, 2006: 22). Redewendungen, dass Worte verletzen oder „wie ein Schlag ins Gesicht" sind, suggerieren dabei eine somatische Dimension des durch Sprache erzeugten Schmerzes (ebd.: 14). Hate Speech sollte demnach nicht durch eine Reihe von Folgen definiert werden, denn sie beschreibt „keine Verletzung und ruft auch keine Verletzung als Folge hervor; vielmehr ist hate speech in der Äußerung selbst die Ausführung der Verletzung" (ebd.: 36).

Eine Fehleinschätzung kommt auch dann zustande, wenn alle Äußerungen, die Schimpfworte enthalten, als Hate Speech erachtet werden. Dieser vermeintliche Zusammenhang bildet aber Grundlage textbasierter Detektierungsalgorithmen. Unter anderem suchen diese automatisiert nach Ethnophaulismen, also herabwürdigenden Bezeichnungen ethnischer Gruppen (Unger, 2013: 280). Diesen Begrifflichkeiten ist eine Dehumanisierung der Bezeichneten inhärent, die in der Geschichte der Schmähworte wurzelt. Über ihre beleidigende Kraft bestehe laut Björn Technau (2013: 229) daher Konsens – auch vor ihrer Verwendung in einem konkreten Kontext. Eine direkte Gleichsetzung von Hate Speech und Schimpfworten ist jedoch nicht möglich, denn die Bedeutung von Worten ist grundlegend durch ihren Kontext bestimmt. So führen beispielsweise wissenschaftliche Auseinandersetzungen das Paradoxon von Nennung und Wirkung vor Augen. Eine Verhandlung von Hate Speech-Ausdrücken ohne deren Zitation ist kaum vorstellbar. Ferner sind Satire und Ironie als Kontexte denkbar (Sirsch, 2013: 169), in denen die Aussagekraft anders ist, als beispielsweise in einem Pamphlet.

Hate Speech geht also über die rein lexikalische Ebene hinaus. Die Gleichsetzung einzelner Indikatoren mit dieser Kommunikationsform bildet lediglich einen simplifizierenden, pragmatischen Ansatz, um ein mannigfaltiges Phänomen in großen Datensätzen, wie beispielsweise sozialen Netzwerken, zu identifizieren (Sponholz, 2018: 68). Eben hierin ist der Grund zu sehen,

warum eine Bestimmung von Hate Speech und somit eine Identifikation von Trollen 2.0 auf qualitativer Ebene erfolgen sollte.

Um Hate Speech als Sprache der Trolle 2.0 zu definieren, lässt sich zusammenfassend folgende Definition festhalten: Hate Speech bezeichnet die öffentliche (Unger, 2013: 257) und intentionale (Sponholz, 2018: 43) Degradierung einer simplifiziert pauschalisierten Kategorie von Menschen (Frischlich et. al., 2017: 71). In bewusster Abwertung (Sponholz, 2018: 60) der vermeintlichen Gruppe oder ihrer Mitglieder werden exkludierende Gegensätze beschworen (Gagliardone et. al., 2015: 11), die der Ausgrenzung und eigenen Machtdemonstration dienen (Schmitt, 2017: 52 f.). Während die Ausführung physischer Gewalt keine Kausalfolge bildet (Brings-Wiesen, 2017: 35), wird, gegebenenfalls unter Verwendung von Schimpfworten, die Verharmlosung von Gewalttaten gegen die Fremdgruppe propagiert (Sponholz, 2018: 43).

Befähigungsfaktoren für Hate Speech in sozialen Netzwerken

Bei der Erforschung von Trollen 2.0 und ihrer Sprache muss dem Umstand Rechnung getragen werden, dass Hate Speech im digitalen Kontext Besonderheiten aufweist, da soziale Netzwerke spezifische Rahmenbedingungen und Befähigungsfaktoren bieten. So hat der Code einer Webseite entscheidenden Einfluss darauf, wie kommuniziert und welche Kommunikation zugelassen wird. Folglich sind Plattformen nicht neutral, denn in ihrer Architektur spiegeln sich die Wertvorstellungen ihrer Programmierer, die den Rahmen definieren, innerhalb dessen sich Nutzer*innen bewegen (Brodnig, 2013: 33).

Zudem konstituiert die Softwarearchitektur die Aufbereitung und Anordnung der Beiträge. Das zentrale, reglementierende Werkzeug hierfür sind die integrierten Evaluationsmechanismen. Die *Like-* respektive *Dislike*-Funktion bildet hierbei die wohl gebräuchlichste Ausformung. Auf Grundlage dieser binären Bewertung entsteht eine Ökonomie der Sichtbarkeit, in der jene Beiträge, die viel Interaktion aufweisen, in der Anordnung nach oben rücken.

Provozierende und diffamierende Inhalte können von dieser Logik profitieren. *Hate Speech* trifft zwar keineswegs ausschließlich auf Zustimmung. Allerdings bezieht sich die algorithmische Wertung als relevant nicht nur auf *Gefällt mir*-Angaben, sondern auf Interaktionen allgemein. In der Folge erlangen Streitthemen besonders hohe Sichtbarkeit und können eine Monopolstellung in den Kommentaren einnehmen. Eine derartige Verzerrung des öffentlichen Diskurses ist besonders dann problematisch, wenn die algorithmische Aufbereitung als Konsens aufgefasst wird.

Katarina Stanoevska-Slabeva (2008: 23) beispielsweise beschreibt die Bewertungsmechanismen als Möglichkeit zur „schnelle[n] Verdichtung und Vernetzung von individuellen Meinungen zu einem kollektiven Meinungspool, der die Meinung der Mehrheit widerspiegelt." Eine repräsentative Aussagekraft erhielten Likes jedoch nur, wenn tatsächlich alle Nutzer*innen eine Wertung abgeben. Unklar nämlich ist, ob Rezipient*innen, die sich an einem In-

halt stören, ebenso aktiv reagieren wie jene, die dem Beitrag zugetan oder ihm gegenüber neutral gestimmt sind. Weiterhin sind Webseitenbesucher*innen ohne Benutzerkonto aus der Evaluation zumeist ausgeschlossen.

Werden jedoch nur die Befähigungsfaktoren der Software betrachtet, droht eine technikdeterministische Perspektive. In diesem Sinne gilt es auch die soziotechnische Seite zu betrachten, also die Aneignung der Strukturen durch die Nutzer*innen. In der Literatur wird diesbezüglich als stimulierender Faktor neben sozialen und kulturellen Aspekten die gefühlte Anonymität im Internet behandelt.

Dieser Aspekt umfasst verschiedene Dimensionen. Einerseits kann in Abwesenheit eines Moderators der Eindruck eines rechtsfreien Raums und einer mangelnden Rechenschaftspflicht entstehen (Fichman / Sanfilippo, 2016: 49). Andererseits trägt auch die Mittelbarkeit des Internets zur gefühlten Anonymität bei. Joachim Knape (2012: 102 f.) spricht von einer tertiärmedialen Kommunikation mit einer Erfahrungsbeschränkung auf einen einzigen Sinn – den visuellen. Soziale Hinweisreize können angesichts softwarearchitektonischer Limitierung nicht vermittelt werden. Ein evidentes Beispiel bilden Mimik und Gestik, deren Fehlen ansatzwiese durch Smileys und Emojis kompensiert wird.

Bekräftigt wird das Gefühl der Anonymität zusätzlich durch ein Phänomen der Gruppendynamik, das in der Psychologie als Deindividuation bezeichnet wird. Dabei nimmt ein Individuum, das sich in einer Gruppe befindet, sich selbst und andere nicht als individuelle Identität wahr (Kaspar, 2017: 64). Zu beobachten ist dieses Phänomen nicht nur in digitalen Kontexten, sondern auch in Massenausschreitungen (Brodnig, 2013: 87). Anonymität führe dabei laut Brodnig (ebd.: 87) „gewissermaßen zu einer Verringerung der Selbstwahrnehmung. Wenn man sich als einer von vielen und damit unerkannt fühlt, sinkt die Angst vor Sanktionen, ebenso das Verantwortungsbewusstsein und die Schuldgefühle."

Die Summe dieser Umstände kann gefühlte Anonymität in einer computervermittelten Kommunikation erhöhen (Kaspar, 2017: 63). Je größer die gefühlte Distanz zu anderen Gesprächsteilnehmern, umso geringer ist die Hemmschwelle, diese zu attackieren und zu diskreditieren. Bezüglich verbaler Enthemmung scheinen Menschen in einer Onlineumgebung besonders leicht beeinflussbar (Cheng et. al., 2017: 2). Wird die Onlinepersönlichkeit gedanklich vom realen Ich getrennt, offenbaren Menschen im Internet teils ein antisoziales Verhalten, das sie offline nicht zeigen (Fichman / Sanfilippo, 2016: 66).

Die Gefahr von Hate Speech in sozialen Netzwerken ist vor allem darin zu sehen, dass sie sich als Umgangston etabliert. Lina Woolf und Michael Hulsizer (2004: 40) schreiben in diesem Kontext: „Hate does not exist in a vacuum. Rather, hate is learned, often from one's family, but also through the groups that one joins." Ähnlich wie beim kindlichen Spracherwerb werden in Onlinegemeinschaften sprachliche Umgangsformen tradiert. Communities bilden beispielsweise interne Slangbegriffe oder Schreibweisen aus. Somit

kann in einer Onlineumgebung, in der Hate Speech hohe Sichtbarkeit erfährt, regelrecht zum Hass erzogen werden (Meibauer, 2013: 5), wodurch sich der Sprachstil der Hassenden in einer Art Ansteckungseffekt konventionalisiert.

Eine amerikanische Studie der Universitäten Stanford und Cornell aus dem Jahr 2017 konnte hierzu nachweisen, dass Nutzer*innen in einer Onlineumgebung, in der bereits ein rauer Umgangston herrscht, dazu neigen, selbst auch zu trollen. Die Untersuchung wies in einer Diskussion mit Trollbeiträgen eine erhöhte Wahrscheinlichkeit nach, dass neue Besucher*innen das Verhalten adaptieren (Cheng et. al., 2017: 1). Die Forscher bestätigen damit, dass neben dem Diskussionsthema auch die Anzahl und Anordnung von Trollkommentaren Einfluss ausübt (ebd.: 8). Daher kommen sie zu dem Schluss, dass sich auch gewöhnliche Nutzer*innen, die zuvor nicht durch Trollverhalten aufgefallen sind, unter entsprechenden Umständen wie Trolle verhalten (ebd.: 1).

Eine exemplarische Analyse: Fremdenhass ohne Ethnophaulismen auf *Facebook*

Aufgrund des zuvor beschriebenen Ansteckungseffekts, sind Trolle 2.0 im Kontext politischer Reizthemen mit bestimmter Wahrscheinlichkeit zu erwarten. Ein Beispiel hierfür bildet ein Interview mit Jörg Radek, dem Bundesvorstand der Gewerkschaft der Polizei, das am 30. Juli 2019 auf der Facebookseite des *ZDF Morgenmagazins* geteilt wurde.[3] Der zugehörige Post lautet:

> In der Diskussion um mehr Sicherheit an Bahnsteigen nach der tödlichen Attacke in Frankfurt ruft Jörg Radek, Gewerkschaft der Polizei – GdP Bundesvorstand, zu Besonnenheit auf, um Nachahmungstäter nicht zu provozieren. Prävention sei bei 5600 Bahnhöfen und mehr als zwei Milliarden Reisenden im Jahr nicht möglich. (@morgenmagazin, 2019)

Das Interview steht in Zusammenhang mit einem Vorfall am Frankfurter Hauptbahnhof einen Tag zuvor, bei dem ein Achtjähriger und dessen Mutter vor einen einfahrenden ICE gestoßen wurden. Als Tatverdächtiger galt ein in der Schweiz lebender Eritreer (Davydov und Iskandar, 2019).

In den Kommentaren wird vermehrt der Vorwurf erhoben, die Politik sehe bei derartigen Verbrechen nur zu, ermahne zur Diplomatie, aber unternehme keine präventiven Schritte (Abb. 1).

Abb. 1: Vorwurf mangelnder politischer Intervention

3 Alle folgenden Kommentare in diesem Kapitel sind Reaktionen auf diesen Post.

Als Lösungsvorschlag wird beispielsweise eingebracht, die Gesetzgebung zu ändern, um härter gegen Täter vorgehen zu können (Abb. 2). Derartige Forderungen werden teils mit rechtem Vokabular wie „Altparteien“ unterlegt (Abb. 3).

Abb. 2: Forderung, gesetzliche Regulierungen zu überarbeiten.

Abb. 3: Verwendung rechten Vokabulars.

Immer wieder wird in den Kommentaren die Jahreszahl 2015 genannt, mit der Begründung, dass Probleme in diesem Ausmaß zuvor nicht existierten (Abb. 4) und bestimmte Sicherheitsmaßnahmen nicht nötig gewesen seien (Abb. 5). In Andeutungen wird damit auf den Umstand Bezug genommen, dass 2015 die Flüchtlingskrise in Europa ihren Höhepunkt erreichte.

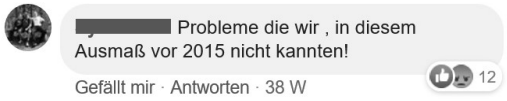

Abb. 4: Benennung des Jahres 2015 als Ursprung nicht näher präzisierter Probleme.

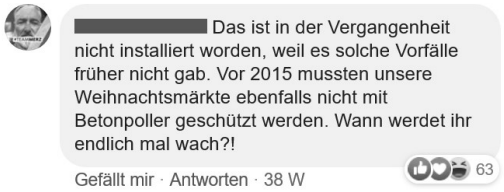

Abb. 5: Bezug auf Sicherheitsmaßnahmen, die hier ab dem Jahr 2015 wahrgenommen werden.

In diesen zuvor genannten Kommentaren – besonders in Abb. 2 – wird eine pauschalisierende Kategorisierung einer Personengruppe zu Tätern bereits tangiert. Es handelt sich um einen klaren *Othering*-Prozess (*wir* vs. *sie*), wenngleich dieser ohne Ethnophaulismen erfolgt. Dennoch vermitteln die nebulösen Andeutungen – mit Bezügen auf politische Diskurse – Abwertung und Ausgrenzung. Simplifizierend werden dabei der Kategorie Ausländer, auf die sich hier bezogen wird, bestimmte Verhaltensweisen zugeschrieben, wie auf der Straße zu „hocken“, zu „saufen“ und zu „pöbeln“ (Abb. 6).

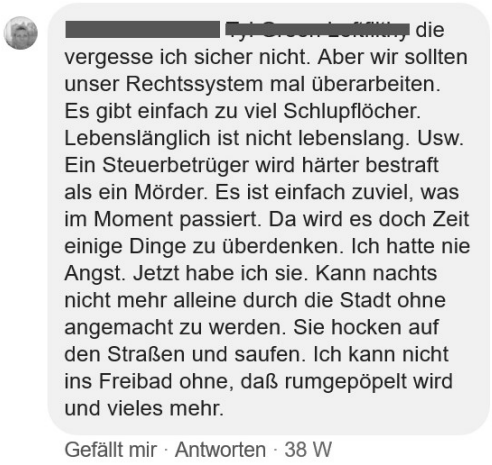

Abb. 6: Kritik am Rechtssystem und Zuschreibung von Verhaltensweisen.

Darüber hinaus wird die sprachlich etablierte Fremdgruppe kriminalisiert (Abb. 7) und als Bedrohung benannt, die die Schreibenden zu vernichten droht (Abb. 8) – erneut eine exkludierende Gegenüberstellung der *Innergroup* und einer äußeren Gefahr.

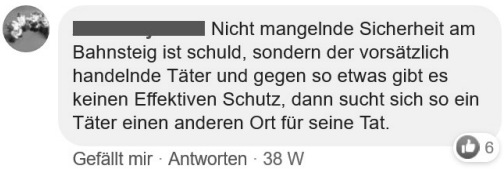

Abb. 7: Schuldzuweisung und Kriminalisierung.

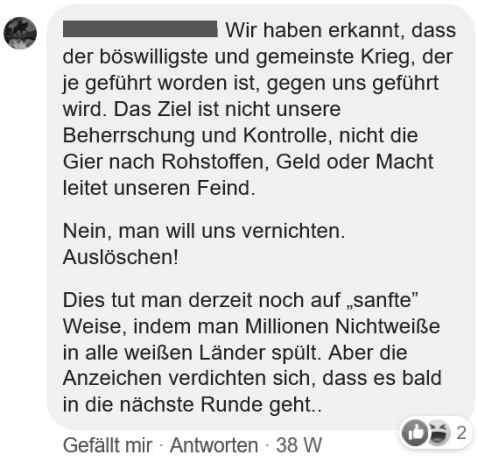

Abb. 8: Konstruktion der Fremdgruppe als Bedrohung und Nennung eines exkludierenden Gegensatzes als Nichtweiße.

In Reaktion auf die konstruierte Bedrohung werden sowohl Handlungsempfehlungen auf politischer Ebene zum Ausdruck gebracht (Abb. 9 und 10) – vornehmlich Abschottung und Ausweisungen – als auch konkrete alltagspraktische Handlungsweisen angeraten (Abb. 11). Abb. 11 ist dabei eine Empfehlung, die diskriminierende Strukturen im alltäglichen Miteinander stärken kann und die zugleich existierenden Rassismus verschriftlicht.

Abb. 9: Forderung nach Grenzschließung.

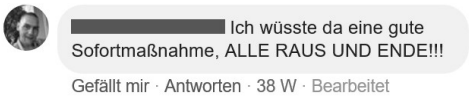

Abb. 10: Forderung nach Ausweisungen.

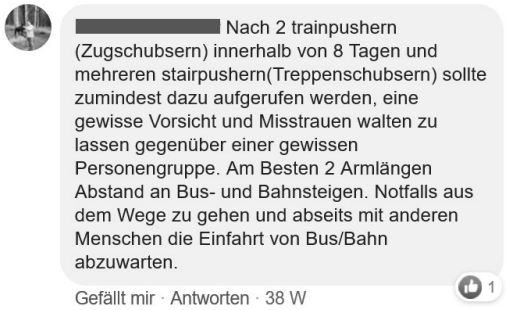

Abb. 11: Aufruf zu Vorsicht und Misstrauen sowie Anpassung des eigenen Verhaltens.

Nicht unerwähnt darf in diesem Beispiel die ausgeprägte Form der Gegenrede bleiben. Immer wieder werden Versuche unternommen, die geäußerten Pauschalisierungen zu relativieren, indem beispielsweise darauf verwiesen wird, dass auch in der *Innergroup* Gewaltbereitschaft existiert (Abb. 12) oder sachlich die Rechtslage erklärt wird (Abb. 13).

Abb. 12: Relativierung der Gewaltbereitschaft unter Migranten.

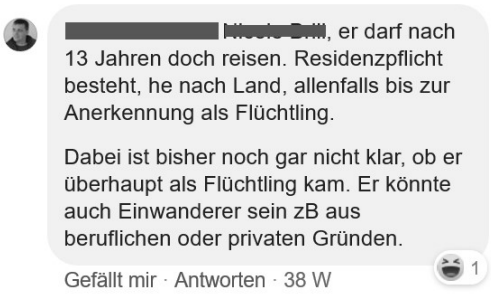

Abb. 13: Darlegung der Rechtslage und Abwägung der Informationen.

Allerdings kommt auch innerhalb der Gegenrede, mit dem Ziel ausgrenzende Kriminalisierung zu unterbinden, gewaltvolle Sprache zum Ausdruck, wie das knappe Beispiel dieses Diskussionszweigs verdeutlicht (Abb. 14).

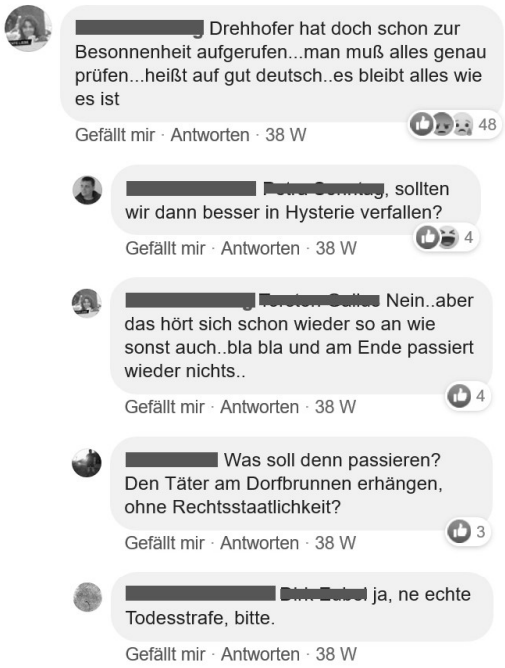

Abb. 14: Diskussionszweig mit Gegenrede unter Verwendung gewaltvoller Sprache.

Abschließend lassen sich bezüglich der hier beispielhaft gewählten Diskussion Besonderheiten im Vergleich zu anderen Plattformen feststellen. So nutzen die Kommentierenden Klarnamen.[4] Sie stehen also mit ihrem Namen zu ausgrenzendem Hass. Die Verwendung von Klarnamen versteht sich nicht

4 Es kann nicht uneingeschränkt davon ausgegangen werden, dass die Profilnamen mit den Personalausweisen übereinstimmen, dennoch ist für einen Großteil der Nutzer*innen zu vermuten, dass es sich um die echten Namen handelt.

als Widerspruch zum Kriterium der gefühlten Anonymität von Trollen 2.0. Wie dargelegt, reicht soziale Anonymität und die gefühlte Distanz zur diffamierten Fremdengruppe aus, um sprachliche Enthemmungen zu fördern. Erwähnt seien auch die auffallend wenigen Rechtschreibfehler. Es ist daher davon auszugehen, dass es sich nicht um affektive Gefühlsausbrüche, sondern gefestigte Überzeugungen handelt – inklusive der Überzeugung, Hate Speech stelle ein probates Mittel dar, um Meinungen zu formulieren.

In exemplarischer Untersuchung der Kommentare zu diesem *Facebook*post wird deutlich, wie *Hate Speech* über die lexikalische Ebene hinausgeht. Würde sie nur über die Verwendung von Schimpfworten definiert, bliebe dieser Fall unentdeckt. Doch eine öffentliche, intentionale Degradierung einer Fremdgruppe mittels abwertender Pauschalisierung wird auch dann erzielt, wenn beispielsweise von einer „gewissen Personengruppe" gesprochen wird von der zwei Armlängen Abstand gehalten werden sollte (Abb. 11).

Wege zur Reaktivierung des demokratisierenden Potenzials sozialer Netzwerke

Obgleich Diensteanbieter im Web 2.0 gemeinsame Merkmale teilen und große Überschneidungen bei den algorithmischen Befähigungsfaktoren aufweisen, lassen sich Abweichungen in der Nutzungssituation feststellen. Um zu eruieren, in welchem Ausmaß Trolle 2.0 Einfluss auf die Umgangsweisen ausüben, muss in allen Bereichen Forschung betrieben werden: über verschiedene Plattformen und unterschiedliche Themenbereiche hinweg. Schließlich finden sich Trolle 2.0 in praktisch jedem Winkel des Internets. Sie sind eine Art Grundrauschen der virtuellen Welt (Glück, 2013: 12) und lassen sich erst durch ihre Benennung erforschen und schließlich bekämpfen. Die größte Gefahr ihres massenhaften Auftretens liegt darin, dass öffentlich sichtbare, scheinbar akzeptierte Hate Speech in sozialen Netzwerken dazu führen kann, dass der aggressive Duktus als angemessener Umgangston aufgefasst wird. Eine Entwicklung, die es zu unterbinden gilt.

Dass eine sprachlich vollzogene Dehumanisierung die Hemmschwelle der Gewaltanwendung senken *kann*, belegt historisch eindrücklich die Degradierung und systematische Vernichtung von Minderheiten im Dritten Reich. Erneut muss an dieser Stelle jedoch daran erinnert werden, dass physische Gewalt keine Kausalfolge von Hate Speech darstellt. Trotzdem kann eine hohe Sichtbarkeit von Hate Speech in sozialen Netzwerken auf gesellschaftlicher Ebene schwerwiegende Folgen haben: Wenn sich Diffamierungen normalisieren, wird Xenophobie unter dem Deckmantel der eigenen Meinung verbreitet und nicht mehr als Grenzüberschreitung wahrgenommen. So wird der *modus operandi* einzelner lauter Stimmen zum *modus vivendi* der Internetgemeinschaft. Damit wird nicht nur gruppenbezogene Menschenfeindlichkeit gestärkt, sondern die Chancengleichheit verletzt, da Hate Speech eine strukturelle Benachteiligung Einzelner oder Gruppen bewirkt (Unger, 2013: 265). Durch

sogenanntes *Silencing* verfälschen Trolle 2.0 damit die Repräsentativität der digitalen Öffentlichkeit, denn die Angst vor Angriffen kann zur freiwilligen Selbstzensur führen. Ziehen sich Minoritäten aber aus sozialen Netzwerken zurück, bleiben ihre Stimmen und Ansichten in der zivilen Sphäre aus. Dies ist als Gefahr für die Meinungsfreiheit zu werten (Fleischhack, 2017: 26).

Gleichzeitig muss sogenanntes *Overblocking* vor dem Hintergrund der utopischen Potenziale des Social Web vermieden werden. Denn das Internet lebt vom Austausch über Grenzen hinweg. In seiner Unabhängigkeitserklärung des Cyberspace machte John Perry Barlow in den 1990er Jahren seinen Wunsch deutlich, einen Ort zu schaffen, an dem Nutzer*innen ihre Überzeugungen zum Ausdruck bringen können. Selbst dann, wenn sie nicht mehrheitskonform sind (Barlow, 1996). Tatsächlich bildet das Internet grundsätzlich ein ebensolches Diskussionsforum. Die Redensart „On the internet, nobody knows you're a dog", die eigentlich vor falschen Identitäten warnen soll, weist genau auf diesen Kerngedanken hin: Online existieren Argumente ohne Bias aufgrund von Alter, Geschlecht oder anderen äußeren Erscheinungen (Brodnig, 2013: 69).

Durch die Eindämmung von Hate Speech kann dieses Potenzial einer idealen Plattform für politischen Austausch reaktiviert werden. Ohne Sorge vor Sanktionen (sei es staatliche Zensur oder ein Trolle 2.0-geführter Shitstorm) können hier Meinungen geäußert werden. Auch wenn diese schockieren oder stören. Schließlich sind auch diese Haltungen dem demokratischen Prozess dienlich. So muss der Schutz *aller* Haltungen – auch jener die von der Mehrheitsmeinung abweichen – im Vorgehen gegen Hate Speech als Maxime gehandelt werden. Ressentiments, Hass und Feindseligkeiten bilden dabei jedoch keine zu schützende Kategorie. Hate Speech ist keine Meinung und keine Haltung Andersdenkender sondern Mittel, um zu dehumanisieren, zu exkludieren und einzuschüchtern.

Die größte Herausforderung bildet dabei die Kontextabhängigkeit. Diese erfordert mühsame Abwägungen in Einzelfallentscheidungen, um Vorfälle einwandfrei bewerten zu können und beispielsweise ironisch formulierte Kritik nicht voreilig zu zensieren. Neben Gerichten und Diensteanbietern können allen voran Communitymitglieder die zukünftige Ausgestaltung des Social Web mitbestimmen. Denn in Anbetracht der Tatsache, dass Trolle 2.0 als situationsbedingtes Phänomen beschrieben werden können, ist ein Umfeld denkbar, in dem der Ansteckungseffekt negativer Kommentare (Cheng et al., 2017) umgekehrt wird. Ein weiterer evolutionärer Schritt sozusagen, zu einem Trolltypus, der Hate Speech als *modus operandi* abstreift und in provokativen Äußerungen proaktiv Denkanstöße auslöst, die demokratische Pluralität an Stimmen befördert und dabei durchaus auch irritiert (Glück, 2013: 42) – schließlich wird gerade im Moment der Unordnung die Ordnung reflektiert und gefestigt.

Literaturverzeichnis

Barlow, John Perry (1996): „A Declaration of the Independence of Cyberspace", in: *Electronic Frontier Foundation* / https://www.eff.org/de/cyberspace-independence.

Bishop, Jonathan (2014): „Representations of ‚trolls' in mass media communication: a review of media-texts and moral panics relating to ‚internet trolling'", in: *International Journal of Web Based Communities*, Vol. 10(1), S. 7–24.

Brings-Wiesen, Tobias (2017): „Das Phänomen der ‚Online Hate Speech' aus juristischer Perspektive", in: Kaspar, Kai / Gräßer, Lars / Riffi, Aycha (Hrsg.): *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, Düsseldorf/München: kopaed, S. 35–48.

Brodnig, Ingrid (2013): *Der unsichtbare Mensch. Wie die Anonymität im Internet unsere Gesellschaft verändert*, Wien: Czernin Verlag.

Butler, Judith (2006): *Haß spricht. Zur Politik des Performativen*, Berlin: Suhrkamp.

Cheng, Justin et. al. (2017): „Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions", in: *Anyone* / https://files.clr3.com/papers/2017_anyone.pdf.

Davydov, Alexander / Iskandar Katharina (2019): „Achtjähriger vor ICE gestossen. Innenminister Seehofer unterbricht Urlaub", in: *Frankfurter Allgemeine Zeitung* / https://www.faz.net/aktuell/rhein-main/frankfurt-hauptbahnhof-mann-stoesst-kind-und-mutter-vor-ice-6307800.html.

Delgado, Richard / Stefancic, Jean (2004): *Understanding words that wound*, Boulder: Westview Press.

Ebner, Julia (2018): *Wut. Was Islamisten und Rechtsextreme mit uns machen*, Darmstadt: Theiss.

Fichman, Pnina / Sanfilippo Madelyn (2016): *Online Trolling and Its Perpetrators. Under the Cyberbridge*, Lanham [u.a.]: Rowman / Littlefield.

Fleischhack, Julia (2017): „Der ‚Hass' der vielen Formen", in: Kaspar, Kai / Gräßer, Lars / Riffi, Aycha (Hrsg.): *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, Düsseldorf/München: kopaed, S. 23–28.

Frischlich, Lena et. al. (2017): „Unmenschlicher Hass: Die Rolle von Empfehlungsalgorithmen und Social Bots für die Verbreitung von Cyberhate", in: Kaspar, Kai / Gräßer, Lars / Riffi, Aycha (Hrsg.): *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, Düsseldorf/München: kopaed, S. 71–79.

Gagliardone, Iginio et. al. (2015): *Countering Online Hate Speech*, Paris: UNESCO Publishing.

Glück, Alexander (2013): *Handbuch für den Forentroll*, St. Ingbert: Röhrig Universitätsverlag.

John, Nicholas A. (2017): *The Age of Sharing*, Cambridge: Polity Press.

Karppi, Tero (2013): „Change name to No One. Like people's status' Facebook Trolling and Managing Online Personas!, in: *The Fibreculture Journal. Digital Media + Networks + Transdisciplinary Critique*, Issue 22, S. 278–300.

Kaspar, Kai (2017): „Hassreden im Internet – Ein besonderes Phänomen computervermittelter Kommunikation?", in: Kaspar, Kai / Gräßer, Lars / Riffi, Aycha (Hrsg.): *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, Düsseldorf/München: kopaed, S. 63–70.

Knape, Joachim (2012): *Was ist Rhetorik?*, Stuttgart: Reclam.

Ley, Hannes (2018): *#ichbinhier. Zusammen gegen Fake News und Hass im Netz*, Köln: DuMont.

Meibauer, Jörg (2013): „Hassrede – von der Sprache zur Politik", in: ders. (Hrsg.): *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, Gießen: Universitätsbibliothek, S. 1–16.

Milner, Ryan M. (2013): „Hacking the Social: Internet Memes, Identity Antagonsim, and the Logic of Lulz", in: *The Fibreculture Journal. Digital Media + Networks + Transdisciplinary Critique*, Issue 22, S. 62–92.

Pohjonen, Matti / Udupa, Shanana (2017): „Extreme Speech Online: An Anthropological Critique of Hate Speech Debates", in: *International Journal of Communication*, 11, S. 1173–1191.

Schaeffer, Ute (2018): *Fake statt Fakt. Wie Populisten, Bots und Trolle unsere Demokratie angreifen*, München: dtv.

Schmitt, Josephine B. (2017): „Online Hate Speech: Definition und Verbreitungsmotivationen aus psychologischer Perspektive", in: Kaspar, Kai / Gräßer, Lars / Riffi, Aycha (Hrsg.): *Online Hate Speech. Perspektiven auf eine neue Form des Hasses*, Düsseldorf/München: kopaed, S. 51–56.

Simek, Rudolf (2018): *Trolle: Ihre Geschichte von der nordischen Mythologie bis zum Internet*, Köln [u.a.]: Böhlau Verlag.

Sirsch, Jürgen (2013): „Die Regulierung von Hassrede in liberalen Demokratien", in: Meibauer, Jörg (Hrsg.): *Hassrede / Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, Gießen: Universitätsbibliothek, S. 165–193.

Sponholz, Liriam (2018): *Hate Speech in den Massenmedien. Theoretische Grundlagen und empirische Umsetzungen*, Wiesbaden: Springer VS.

Stanoevska-Slabeva, Katarina (2008): „Web 2.0 – Grundlagen, Auswirkungen und zukünftige Trends", in: Meckel, Miriam / Stanoevska-Slabeva, Katarina (Hrsg.): *Web 2.0. Die nächste Generation Internet*, Baden-Baden: Nomos, S. 13–38.

Technau, Björn (2013): „Sprachreflexion über politisch inkorrekte Wörter: Eine konversationsanalytische Studie", in: Meibauer, Jörg (Hrsg.): *Hassrede / Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, Gießen: Universitätsbibliothek, S. 223–256.

Unger, Doris (2013): „Kriterien zur Einschränkung von hate speech: Inhalt, Kosten oder Wertigkeit von Äußerungen?", in: Meibauer, Jörg (Hrsg.): *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, Gießen: Universitätsbibliothek, S. 257–285.

Whillock, Rita Kirk / Slayden, David (1995): „Introduction", in: dies. (Hrsg.): *HATE SPEECH*, Thousand Oaks [u.a.]: Sage Publications, S. ix–xvi.

Wilson, Jason et. al. (2013): „Trolls and the Negative Space of the Internet. Editorial", in: *Directory of Open Access Journals* / https://doaj.org/article/9cecebbcf5864afc900d60770fe7b1a9.

Woolf, Lina M. / Hulsizer, Michael R. (2004): „Hate groups for dummies: How to build a successful hate group", in: *Humanity and Society*, Vol. 28(1), S. 40–62.

ZDF Morgenmagazin [@morgenmagazin] (2019): „Facebook-Post", in: *Facebook* / https://www.facebook.com/morgenmagazin/videos/700680047026997/l.