Gutachten zum Phänomen des digitalen Hasses und der digitalen Hetze in Niedersachsen:

Bedingungen, Erscheinungsformen und Ansätze der Prävention und Intervention

Angefertigt von:
Samuel Tomczyk

Unter Mitarbeit von:

Johanna Brüser,
Olivia Feltz, Denise Jais
und Maxi Wiesemüller

gefördert durch
Landesprogramm
für Demokratie
und Menschenrechte

Inhaltsverzeichnis

Einleitung 3

Definition des Phänomenbereichs 5

Perspektive auf Betroffene 8

Perspektive auf Ausübende 13

Perspektive auf Beobachtende 18

Die Rolle von Medium und Umwelt 21

Theoretische Zugänge zur Entstehung und Prävention von Hass im Netz 28

Verhältnis- und verhaltensbasierte Ansätze der Prävention und Intervention 33

Die Situation in Niedersachsen 45

Implikationen für die Forschung und Praxis 47

Angaben zur Positionalität 55

Literaturverzeichnis 56

Einleitung

Die Phänomene digitaler Hass und digitale Hetze sind seit einiger Zeit Teil unseres Alltags, insbesondere für Kinder, Jugendliche und junge Erwachsene – rund 92 % der jungen Menschen bis 25 Jahre, die überwiegend online aktiv sind, haben bereits direkt oder indirekt (z. B. als Zeug*innen[1]) Erfahrungen mit digitalem Hass/digitaler Hetze gemacht, indem sie abwertende, entwürdigende, auf Einschüchterung zielende oder verhetzende Onlineinhalte produziert, verbreitet oder rezipiert haben (Landesmedienanstalt NRW, 2022; https://www.medienanstalt-nrw.de/hass.html).

Ein prominentes Beispiel für digitalen Hass und digitale Hetze ist Hassrede/*Hate speech*, die sich auf Kommunikationsinhalte (z. B. Worte, Bilder, Videos) bezieht, welche sich auf Basis gruppenbezogener und identitätsrelevanter Merkmale (z. B. Geschlecht, Religion, Ethnie, sexuelle Orientierung) gegen Personen richten und diese diffamieren, bedrohen oder schädigen sollen (Chetty & Alathur, 2018; Kansok-Dusche et al., 2023; Sponholz, 2020, 2021). Hassrede basiert auf der Intention, andere zu verletzen und kann auf individueller (z. B. Stress, Wut, Trauer), sozialer (z. B. Anfeindung) und gesellschaftlicher (z. B. Ausgrenzung) Ebene Folgen haben.

Ein weiterer Begriff im Themenfeld digitaler Hass bzw. digitale Hetze ist „digitale Gewalt“, die das Ziel hat, anderen Personen in digitalen Räumen oder über virtuelle Kanäle oder Mittel, Schaden zuzufügen, sie dem eigenen Willen zu unterwerfen oder als Reaktion auf erfahrene (digitale) Gewalt,

[1] Der Text ist in geschlechtergerechter Sprache gestaltet und nutzt den Genderstern, falls eine alternative Formulierung sehr umständlich ist.

Gegen-Gewalt einzusetzen (Sponholz, 2020). Damit stellt sie eine Sonderform dar, die aber strukturelle Gemeinsamkeiten mit anderen Bereichen besitzt.

Als ein Sammelbegriff für diese Phänomene kann daher „Hass im Netz“ gelten, wie ihn etwa das Kompetenznetzwerk gegen Hass im Netz gebraucht (https://kompetenznetzwerk-hass-im-netz.de/). Es handelt sich dabei nicht um einen eindeutigen, juristisch definierten Begriff. Da grundsätzlich die Meinungsfreiheit (Art. 5, Abs. 1 GG) – auch im digitalen Raum – als schützenswert gilt, gelten Äußerungen und Handlungen, die unter „Hass im Netz“ gefasst werden können, nicht zwingend als Straftatbestand. Sofern eine digitale Handlung allerdings andere Rechte einschränkt oder bedroht, z. B. Menschenwürde oder Persönlichkeitsrechte, kann es strafrechtlich relevant werden. Dazu zählen etwa Beleidigung (§185 StGB), Üble Nachrede und Verleumdung (§§ 186 und 187 StGB), Nötigung (§240 StGB), Bedrohung (§241 StGB), Belohnung und Billigung von Straftaten ($140 StGB), Nachstellung/Stalking (§238 StGB), Verbreitung pornographischer Inhalte (z. B. an Minderjährige; §184 StGB), Gefährdendes Verbreiten personenbezogener Daten (z. B. sog. Doxxing, §126a StGB), Volksverhetzung (§130 StGB) oder auch das Verbreiten von Propagandamitteln verfassungswidriger und terroristischer Organisationen (§86 StGB), die als Propaganda gewaltbefürwortende Inhalte, Hass und/oder Hetze transportieren.

In der Auseinandersetzung mit dem Themenfeld besteht daher stets ein Spannungsfeld von Meinungsfreiheit, straffreier und strafrechtlicher relevanter Kommunikation in digitalen Räumen. Die große Bandbreite an potenziell

strafrechtlich relevanten Beständen einerseits und der Vielfalt an Erscheinungsformen von Hass im Netz andererseits, auch mit Blick auf das schützenswerte Gut der Meinungsfreiheit, verdeutlicht die Komplexität des Themenfelds. Um auf individueller, sozialer und gesellschaftlicher Ebene einen angemessenen und verantwortungsvollen Umgang zu ermöglichen, ist es erforderlich, sich mit den Grundlagen von digitalem Hass und digitaler Hetze auseinanderzusetzen, ihre Entwicklung und Erscheinungsformen zu betrachten und erfolgsversprechende Ansatzpunkte der Prävention und Intervention zu identifizieren. Dieses Gutachten betrachtet daher zunächst die theoretischen und empirischen Grundlagen zum Thema, mit Blick auf die Situation in Niedersachsen, um davon ausgehend Maßnahmen zu erörtern, die zur Prävention sowie zur Intervention und Reaktion bei Hass im Netz zur Verfügung stehen. Aus der Zusammenschau sollen dann weiterführende Impulse für Forschung und Praxis abgeleitet werden.

Definition des Phänomenbereichs

Neben den eingangs dargestellten, strafrechtlich relevanten Ausdrucksformen von Hass im Netz lassen sich weitere, verwandte Konzepte beschreiben, die ihrerseits diverse Forschungstraditionen besitzen und Implikationen für die Entstehung, Aufrechterhaltung, Prävention und Intervention liefern (für eine Analyse der Forschungstrends zu Online-Hass, siehe Waqas et al., 2019).

Die Konzepte lassen sich entsprechend ihrer Perspektive auf das Medium und Umfeld und die beteiligten Personen (als ausführende, erlebende oder

beobachtende Person von Hass im Netz[2]) kategorisieren und sie differenzieren darüber hinaus z. T. die Form (z. B. Bilder, Videos, Textnachrichten), die Intention und den Inhalt (z. B. ideologische Ausrichtung) der Hassbotschaft (Mohseni, 2023). Gemein ist den Konzepten, dass sie den Umstand beschreiben, dass Personen sich digital vermittelt gegen andere Personen (Individuen oder Gruppen) wenden und diese abwerten, bedrohen, beleidigen oder verletzen. Nicht darunter gefasst werden digitale Angriffe auf nichthumane Sicherheitssysteme oder cyberphysische Systeme ohne zwingende menschliche Einbindung. Das Beobachten (auch: indirektes Erleben) von Hass im Netz ist dabei deutlich häufiger anzutreffen als das Ausüben, gefolgt vom direkten Erleben von Hass im Netz, basierend auf Selbstauskunft von Personen (Henares-Montiel et al., 2022; Kansok-Dusche et al., 2023).

In der Fachliteratur werden Unterschiede zwischen Phänomenen wie Hassrede, digitaler Hetze, Hassbeiträgen/postings und Cyberbullying oder Cyberaggression diskutiert, die Implikationen, etwa für die kommunikationswissenschaftliche Forschung, besitzen (Sponholz, 2021). Dies ist aus Forschungsperspektive notwendig, stellt die Praxis allerdings vor Herausforderungen, da es detailreiches Fachwissen erfordert und impliziert, dass spezifische Zugänge für Prävention und Intervention nötig sein könnten. Bislang liegen für diese Annahme jedoch wenig belastbare Befunde vor, was verständlich ist, da das Feld vergleichsweise jung ist (seit Anfang der 2000er Jahre).

[2] Um der Multidisziplinarität und inhaltlichen Vielfalt des Themenfelds gerecht zu werden, wird bewusst auf Bezeichnungen wie Täter oder Opfer verzichtet, da diese v. a. in der kriminologischen Betrachtung gebräuchlich sind (Dölling, 2012).

Damit erscheinen vor allem Gemeinsamkeiten und übergreifend wirksame Präventions- und Interventionsansätze für die Praxis prioritär, ohne dabei Spezifika aus den Augen zu verlieren und laufend wissenschaftliche Erkenntnisse zu neuen Phänomenen und Ansätzen in die Praxis zu spiegeln. Dieses Argument wird durch aktuelle Forschung unterstützt (z. B. Fulantelli et al., 2022 zum Thema Cyberaggression), da phänomenologisch, ätiologisch und präventions- bzw. interventionsorientiert große Überschneidungen zwischen einigen Konzepten bestehen. Aus diesen Gründen sollen nachfolgend vor allem Gemeinsamkeiten und damit übergreifende Ansatzpunkte herausgearbeitet werden, die mit Blick auf die öffentliche Sicherheit und Gesundheit relevant sind. Zunächst stehen die beteiligten Personen im Fokus.

Zusammenfassung: Definition von Hass im Netz

- „Hass im Netz“ bezeichnet als Sammelbegriff Phänomene wie digitale Gewalt, Hassrede, Cyberbullying oder -aggression
- Die Phänomene unterscheiden sich in ihrem Auftreten und werden z. T. unterschiedlich definiert, besitzen aber wesentliche Gemeinsamkeiten in der Entstehung sowie hinsichtlich zentraler Schutz- und Risikofaktoren
- Phänomene von Hass im Netz bewegen sich zwischen strafrechtlich relevantem Verhalten und freier Meinungsäußerung und müssen deshalb auf Fallbasis betrachtet werden

Perspektive auf Betroffene

Forschung zu Personen, die als Betroffene Hass im Netz erfahren, findet sich vielfach unter dem Schlagwort **Cyberviktimisierung**, in der Tradition der Viktimisierungsforschung (Vranjes et al., 2018). Dem Bereich können Erlebnisse wie Cyberstalking (Kaur et al., 2021), Cyberharassment und Cyberdiskriminierung (Stevens et al., 2021) sowie Cyberbullying/mobbing (Sabella et al., 2013) zugeordnet werden. Sie unterscheiden sich etwa anhand einer Fokussierung auf Einschüchterung oder Abwertung (Harassment bzw. Diskriminierung), Aufdringlichkeit und Grenzverletzung (Stalking) oder Ausnutzung oder Reproduktion bestehender Machtgefälle (z. B. sozialer Status; Mobbing). Alle drei Bereiche stehen bei Betroffenen in Zusammenhang mit eingeschränktem Wohlbefinden und Selbstwertgefühl, Angst- und Stresserleben, posttraumatischer Stresssymptomatik und einer höheren Wahrscheinlichkeit der Entwicklung psychischer Störungen wie Depressionen oder Posttraumatischer Belastungsstörung sowie – insbesondere im Kindes- und Jugendalter – selbstverletzendem Verhalten bis hin zu suizidalen und parasuizidalen Handlungen (Buelga et al., 2022; Jadambaa et al., 2019; John et al., 2018; Li et al., 2022). Die Zusammenhänge sind dabei vergleichbar und z. T. stärker als bei Mobbing, das außerhalb digitaler Räume stattfindet (Henares-Montiel et al., 2022; Petras & Petermann, 2019).

Eine Übersichtsarbeit zum Zusammenhang von Cyberviktimisierung mit chronischen Gesundheitsproblemen (Alhaboby et al., 2019) verweist zudem auf ein höheres Risiko für Cyberviktimisierung von Personen mit chronischer Gesundheitsstörung und beschreibt das Auftreten von unspezifischen

somatischen Symptomen (z. B. Kopfschmerzen, Schlafstörungen) als häufig berichtete Folge von Cyberviktimisierung. Diese gesundheitsrelevanten Folgen verdeutlichen die Bedeutung von Hass im Netz für die Gesundheit der Bevölkerung. Außerdem wird deutlich, dass bereits vulnerable Gruppen (z. B. Personen mit chronischer Gesundheitsstörung) besonders schützenswert sind.

In der epidemiologischen Forschung gelten Aspekte wie eine chronische Erkrankung, aber auch soziodemografische Merkmale wie Geschlecht, Alter, Einkommen, Berufs- oder Bildungsstatus als soziale Determinanten von Gesundheit und werden auch als Dimensionen der Ungleichheit betrachtet (sog. *Health Inequity*, vgl. Karran et al., 2023). Sie bezeichnen Merkmale, die Gruppen von Menschen voneinander unterscheiden und die eine Bedeutung für die Entwicklung, Aufrechterhaltung und Wiederherstellung von Gesundheit haben, z. B. durch ihren Einfluss auf das Gesundheitsverhalten von Personen, die Verfügbarkeit und Inanspruchnahme von Gesundheitsleistungen oder auch die Wirksamkeit von Maßnahmen (z. B. im Sinne der Gendermedizin). Während die Forschung zu medizinischen Phänomenen in diesem Kontext weit fortgeschritten und vergleichsweise strukturiert ist (in der Studie von Karran et al. wurden 200 Primärstudien zum Thema eingeschlossen), gilt dies für andere Bereiche, wie etwa Hass im Netz, in deutlich geringerem Maße. Zumeist werden dort nur einzelne Dimensionen von Ungleichheit untersucht und das Zusammenwirken mehrerer Faktoren wird nicht umfassend berücksichtigt, systematische Analysen relevanter Dimensionen der Ungleichheit stehen aus. Das Betroffensein von Personen mit chronischer Gesundheitsstörung durch Hass im Netz verdeutlicht allerdings die Relevanz

einer solchen übergeordneten Perspektive, die Hass im Netz auch im Kontext von Gesundheit und sozialer Ungleichheit betrachtet.

Ein Blick auf weitere Dimensionen der Ungleichheit – neben der chronischen Gesundheitsstörung – wie Geschlecht, zeigt, dass sich zwar in dieser Gruppe keine bedeutsamen Unterschiede in der Häufigkeit zwischen den Geschlechtern zeigten, psychosoziale und psychosomatische Belastungen als Folge von Viktimisierung aber deutlich häufiger von Frauen bzw. Mädchen berichtet wurden. Sie erleben häufiger Hass im Netz in Zusammenhang mit sexuellen Inhalten und damit vermutlich eine höhere Belastung. Zu klären ist darüber hinaus, ob dies darüber hinaus Zeichen intersektionaler Benachteiligung ist, von der weibliche Personen mit Gesundheitsproblemen besonders betroffen sind, oder ob Belastungen von männlichen Personen unterrepräsentiert werden, bspw. um Männlichkeitsidealen von Unverletzlichkeit und Stärke zu entsprechen, die einem Eingestehen von Belastung und der Inanspruchnahme von Hilfe und Unterstützung entgegenstehen können (vgl. Easton, 2014; Seidler et al., 2016; Wright & Wachs, 2020). Dimensionen der Ungleichheit, die bislang beforscht wurden, sind etwa Alter, Einkommen, Region, ethnische Zugehörigkeit. Überwiegend zeigt sich, dass Personen, die mit Vulnerabilität verbundene Ungleichheit aufweisen, weil sie z. B. auf geringere soziale Macht hinweisen (wie Teil einer ethnischen Minderheit, weibliches Geschlecht, geringeres Einkommen), im Trend etwas häufiger und zumeist stärker von Cyberviktimisierung betroffen sind (z. B. häufiger psychische Störungen entwickeln) (Campbell et al., 2019; Li et al., 2022; Lozano-Blasco et al., 2023; Martinez-Cao et al., 2021; Petras & Petermann, 2019).

Die Wahrscheinlichkeit, von Cyberviktimisierung betroffen zu sein, ist für Personen erhöht, die weitere Vulnerabilität aufweisen, z. B. frühere Viktimisierungserfahrung, geringer Selbstwert oder geringe soziale Anbindung außerhalb digitaler Räume (Marín et al., 2019). Zudem unterscheiden sich die Befunde je nach kulturellem Hintergrund, sodass dies bei der Gestaltung von Maßnahmen zu beachten ist. Dies ist für die Präventionsforschung und -praxis bedeutsam, da z. B. für den deutschsprachigen Raum nicht so viele längsschnittliche Studien und qualitativ hochwertige Evaluationsstudien für Interventionen wie für den internationalen Raum (v. a. USA) vorliegen. Damit ist die Aussagekraft eingeschränkt und die Empfehlungen sind unter Vorbehalt auszusprechen. Insbesondere für die partizipative Entwicklung, Implementation und Evaluation von Interventionen, die relevante Dimensionen der Ungleichheit berücksichtigen, besteht Forschungsbedarf, sodass der Punkt an dieser Stelle nicht vertieft werden kann. Gleichwohl lässt sich aus diesen Befunden ein Hinweis auf die Nützlichkeit zielgruppengerechter Intervention ableiten, vergleichbar mit Ansätzen im Gesundheitswesen (z. B. Braveman, 2006). Die dortige Forschung zeigt, dass das Zuschneiden von Maßnahmen (sog. *Tailoring*) auf relevante Personenmerkmale, wie z. B. Geschlecht, kultureller Hintergrund, Motivation, Ausmaß des Zielverhaltens (etwa Rauchen, Bewegung), gegenüber generischen Maßnahmen, wie z. B. allgemeine Informationen zu Vorteilen des Nichtrauchens, hinsichtlich Wirksamkeit und Nachhaltigkeit der Maßnahmen einen Mehrwert bedeutet (Krebs et al., 2010; Ryan & Lauver, 2002).

Nicht zuletzt aufgrund der gesundheitsbezogenen Konsequenzen von Cyberviktimisierung werden daher zunehmend Forderungen laut, die Hass im

Netz als Public Health-Problem betrachten und entsprechend politisches und gesellschaftliches Handeln fordern, das z. B. über projektbezogene Finanzierung im Bereich Demokratieförderung hinausgeht, und der Komplexität der Thematik gerecht wird (vgl. Jamison et al., 2019; Nguyen, 2023).

Ein beispielhafter Ansatzpunkt für eine solche Arbeit ist die Kombination der Arbeit von Informatik, Sicherheitstechnik und Sozialwissenschaft (z. B. zur Entwicklung und Evaluation von Algorithmen zur Erkennung und Analyse von Hass im Netz mit einem Fokus auf Erklärbarkeit der Prozesse), Demokratie- und Medienbildung (z. B. zur Identifikation demokratieförderlicher und -schädlicher Narrative und Möglichkeiten der (digitalen) Partizipation) und Public Health (z. B. zur Erfassung von Indikatoren der Betroffenheit von Bevölkerungsgruppen und der Ableitung von Empfehlungen zur gezielten Intervention). Da die Entwicklung von Public Health-Maßnahmen häufig langsam vorangeht und ressourcenintensiv ist und somit der Dynamik digitaler Kommunikation bisweilen entgegenstehen kann, ist daher eine engmaschige Zusammenarbeit dieser Bereiche, eingerahmt durch Aufklärung und Bildungsarbeit, zu empfehlen (vgl. Nguyen, 2023).

Zusammenfassung: Perspektive auf Betroffene

- Betroffene von Hass im Netz erleben häufig Angst und Stress, haben ein höheres Risiko, psychische Störungen zu entwickeln und berichten geringeren Selbstwert und geringeres Wohlbefinden
- Personen, die soziale Ungleichheit aufweisen (z. B. als ethnische Minderheit, Person mit chronischer Erkrankung), sind besonders vulnerabel für Hass im Netz und so häufig mehrfach benachteiligt
- Weibliche Personen sind nicht unbedingt häufiger von Cyberhass betroffen, aber berichten i. d. R. stärkere Belastungen
- Aufgrund der klaren gesundheitlichen Assoziationen kann Hass im Netz auch als Public Health-Problem betrachtet werden

Perspektive auf Ausübende

Das Ausüben von Hass im Netz ist Gegenstand multiperspektivischer Forschung, um Risiko- und Schutzfaktoren für das Auftreten und Prozessvariablen im Geschehen zu identifizieren, z. B. für das Ausüben von Cyberbullying oder Hassrede, sodass entsprechende Ansätze der Prävention und Intervention gestaltet werden können. In der Forschung wurde lange ein Fokus auf das Thema Cyberbullying gelegt; der Begriff wurde z. T. auch als Sammelbegriff genutzt, um gegen Personen gerichtetes, digitales Verhalten zu beschreiben, wie digitale Hassrede, Happy Slapping (Filmen und Veröffentlichen körperlicher Gewalt), Impersonation (Auftreten unter dem Namen einer anderen Person), Flaming (starke verbale Auseinandersetzungen), Harassment bzw. Diskriminierung und Beleidigung, Cybergrooming (Anbahnung sexueller Kontakte mit Minderjährigen), Cyberstalking, Cyberbedrohung, Outing bzw. Doxxing, Denigration (Verleumdung, Verbreitung von Gerüchten) und gezielter Ausschluss (z. B. aus Gruppen in sozialen Medien) (Castaño-Pulgarín et al., 2021; Sabella et al., 2013).

Mit der Zeit wurde eine zunehmende Differenzierung dieser Bereiche angeregt, um die Gemeinsamkeiten und Unterschiede zu betrachten, genauer zu erforschen und für die Praxis nutzbar zu machen. Die Abgrenzung und Definition dieser Konzepte (so wie die Diskussion neuartiger Phänomene wie Doxxing) sind weiterhin Gegenstand aktueller Diskussion in der Forschung (vgl. etwa Kansok-Dusche et al., 2023; Sponholz, 2021). So wird etwa diskutiert, was konstituierende Merkmale von Hassrede sind, dazu gehören Spontaneität, Gruppenbezug und Assoziation mit Vorurteilsstrukturen (z. B.

geringschätzige Annahmen über Menschen mit Migrationshintergrund). Diskutiert wird, wie Hassrede im Verhältnis zu Hassbeiträgen (z. B. als einmalige Beiträge) steht und wie genau sie im Vergleich zu anderen Formen von Hass im Netz einzuordnen ist, wie z. B. Cyberbullying, für das – anders als bei Hassrede – häufig gezieltes, planhaftes, personenbezogenes und wiederholtes Agieren als Definitionsmerkmale beschrieben werden (z. B. Abwertung einer Person aufgrund ihrer Kleidung oder ihres Aussehens). Auch im Bereich Cyberbullying existieren diverse Definitionen und Erhebungsmethoden, die für die Konsensbildung und die Evidenzsynthese herausfordernd sind (Sabella et al., 2013).

Wichtig zu betonen ist, dass verschiedene Phänomene von Hass im Netz unabhängig voneinander auftreten und sich gegenseitig verstärken können. Zudem liegen für einige Formen von Hass und Gewalt im Netz wie Happy Slapping deutlich weniger Befunde vor als für andere, da diese Phänomene seltener auftreten, eine höhere Hemmschwelle besitzen (z. B. verglichen mit Flaming) oder sich auf gewisse geographische Räume fokussieren (Ching et al., 2012, 2017). Eingedenk dieser Einschränkungen lassen sich dennoch übergreifend Faktoren identifizieren, die häufig mit dem Ausüben von verschiedenen Ausdrucksformen von Hass im Netz verbunden sind (z. B. Fulantelli et al., 2022), weshalb sie an dieser Stelle unter dem Oberbegriff Cyberhass zusammengeführt werden, der Aspekte der Cyberaggression (z. B. Cyberbullying) sowie der verbalen und nonverbalen Kommunikation (z. B. Hassrede, Flaming) umfasst (Blaya, 2019).

In systematischen Übersichtsarbeiten und Meta-Analysen zu Cyberhass (Chen et al., 2017; Evangelio et al., 2022; Graf et al., 2019; Guo, 2016; Henares-Montiel et al., 2022; Lo Cricchio et al., 2021; Rudnicki et al., 2023; Zych et al., 2019) zeigt sich, dass Personen, die Cyberhass ausüben, tendenziell häufiger männlich und älter sind (wenngleich die Effekte eher klein sind), ausgeprägtes externalisierendes Verhalten zeigen (z. B. sozial auffälliges, aggressives Verhalten) und generell eine hohe Online-Aktivität berichten, insbesondere in sozialen Medien.

Auf Ebene der Persönlichkeit finden sich zudem Zusammenhänge des Ausübens von Cyberhass mit einer höheren Ausprägung von Ärger, Aggression, antisozialen Tendenzen sowie geringer Verträglichkeit und geringer (affektiver) Empathie (Graf et al., 2019). Für politische Einstellungen liegen gemischte Befunde vor, z. B. für rechtsgerichteten Autoritarismus. Im Bereich der politischen Radikalisierung und Extremismusforschung wird Autoritarismus als ein relevanter Einflussfaktor auf Vorurteilsbildung und die Annahme radikaler politischer Ideologien diskutiert (vgl. Cuevas & Dawson, 2021), im Kontext von Hassrede im Internet zeigt sich allerdings auch ein anderes Bild, wonach Autoritarismus mit einer Ablehnung von Hassrede assoziiert sein kann (Bilewicz & Soral, 2020). Der Befund wird damit erklärt, dass die Hassrede als Normverletzung gesehen wird, die sich den an klarer Ordnung ausgerichteten Überzeugungen entgegenstellt. Gleichsam besteht ein positiver Zusammenhang von Autoritarismus mit der Ablehnung von Fremdgruppen und Vorurteilsstrukturen, welche wiederum andere Formen der Ausgrenzung und Diskriminierung wie Mobbing und schädigendes Verhalten (auch außerhalb digitaler Räume) begünstigen.

Andere Einstellungen wie soziale Dominanzorientierung sind positiv mit Cyberhass assoziiert (Bilewicz & Soral, 2020). Der genaue Einfluss politischer Überzeugungen und Werteorientierungen auf die Wahrnehmung, Bewertung und Praxis von Cyberhass ist daher eingehend zu prüfen. Ein weiterer wichtiger Befund ist, dass das Risiko für Cyberhass erhöht ist, wenn Personen selbst Erfahrungen mit Cyberviktimisierung oder Mobbing außerhalb digitaler Räume gemacht haben (als beobachtende, betroffene oder ausführende Person) (z. B. Fulantelli et al., 2022; Jadambaa et al., 2019; Petras & Petermann, 2019).

Insbesondere im Jugendalter zeigt sich, dass das Ausüben von Cyberhass auf psychologischer Ebene mit Prozessen der sozialen Verstärkung, negativen Zukunftsbildern und Moral Disengagement[3] verbunden ist (Bandura, 2011; Lo Cricchio et al., 2021; Nocera et al., 2022; Runions & Bak, 2015; Wachs et al., 2022). So ist die Wahrscheinlichkeit für das Ausüben höher, wenn soziale Verstärkung als Motiv relevant ist, weil z. B. Peers Cyberbullying betreiben oder gutheißen (z. B. durch Likes, Kommentare) oder höhere Popularität als Belohnung antizipiert wird (die aber nicht unbedingt objektiv vorliegen muss) (Walther, 2022; Wegge et al., 2016). Negative Zukunftsbilder, als Gegenentwurf zu positiven Zukunftsbildern, bedeuten eine pessimistische Haltung gegenüber der eigenen Zukunft, der Chancen auf soziale Mobilität und Erreichung eines positiv besetzen Sozialstatus (Miller & Brickman, 2004; Tomczyk et al., 2023). Personen mit negativen Zukunftsbildern haben den Eindruck, dass ihre Handlungen wenig Einfluss

[3] Moral Disengagement bezeichnet die Trennung von eigenen Handlungen und internen moralischen Überzeugungen als Rechtfertigung für nichtnormatives oder deviantes Verhalten und damit sozial unerwünschtes Verhalten (vgl. Bandura, 2011).

auf die Zukunft haben, erleben mehr Frustration und Enttäuschung, und sind daher anfälliger für Disinhibition[4] und motivierter für den Ausdruck negativer Werthaltungen, was den Ausdruck von Cyberhass begünstigt (Casale et al., 2015). Moral Disengagement im Kontext von Cyberhass drückt sich etwa durch Verantwortungsdiffusion (bei sehr vielen, anonym online anwesenden Personen), die vermeintliche Harmlosigkeit der Handlung (wenn Cyberhass z. B. als Ausdruck von Humor bezeichnet wird) und ausbleibende (sichtbare) Konsequenzen des Handelns als Rechtfertigung aus (Lo Cricchio et al., 2021). Moral Disengagement begünstigt daher verschiedene Formen der Beteiligung an Cyberhass, z. B. das Erstellen oder Verbreiten von Hassbotschaften, aber auch das Einsetzen von Gegen-Gewalt und die aggressive Reaktion auf Hass (vgl. auch Flaming) sowie Bystander-Effekte (s. u.) (Runions & Bak, 2015; Wachs et al., 2022). So kann z. B. erklärt werden, warum Betroffene von Cyberhass selbst Cyberhass ausüben können.

Wenngleich die Mehrheit bisheriger Forschung den Kinder- und Jugendbereich fokussiert, zeigen sich für ältere Stichproben vergleichbare Zusammenhänge (z. B. Domínguez-Hernández et al., 2018; Vranjes et al., 2018), was als ein Hinweis auf entwicklungsorientierte Prozesse gelten kann, wie sie etwa im Bereich der Radikalisierung postuliert werden (vgl. Beelmann, 2020) mit Blick auf deviante Sozialentwicklung. Bislang existieren allerdings kaum Längsschnittstudien, um diese Annahmen empirisch zu sichern.

[4] (Online-)Disinhibition meint eine fehlende Zurückhaltung im Ausdruck von sozial unerwünschtem Verhalten, etwa aufgrund des Wegfalls von Barrieren (z. B. wahrgenommene Anonymität), der Erhöhung der Erfolgswahrscheinlichkeit (z. B. durche die Einfachheit der Kommunikation) und der zunehmenden Attraktivität sozial unerwünschten Verhaltens (z. B. durch sog. Echokammern in sozialen Medien) (vgl. Casale et al., 2015).

Zusammenfassung: Perspektive auf Ausübende

- Verschiedene Phänomene von Hass im Netz können unabhängig voneinander auftreten und sich gegenseitig verstärken (so ist z. B. auch eine Eskalation digitaler Aggression möglich)
- Personen, die Hass im Netz ausüben/verbreiten, sind häufiger männlich, älter, zeigen eher externalisierendes Verhalten und berichten hohe Online-Aktivität/Affinität (v. a. bezüglich sozialer Medien)
- Sie erreichen höhere Werte in Tests zu Aggression und antisozialen Tendenzen, Moral Disengagement, sozialer Dominanzorientierung und geringere Werte im Bereich Empathie
- Eigene Erfahrungen (als Betroffene*r) sind ein Risikofaktor für zukünftiges Erleben und Ausüben von Hass im Netz
- Die Zusammenhänge wurden bislang überwiegend im Kindes- und Jugendalter untersucht, finden sich aber auch bei Erwachsenen

Perspektive auf Beobachtende

Da die Wahrnehmung von Cyberhass ohne aktive Beteiligung (sog. *Bystander*) besonders häufig ist, gilt es, diese Gruppe in den Blick zu nehmen, um für sie relevante Ansatzpunkte der Prävention zu identifizieren, damit sie einerseits nicht selbst zu Ausübenden werden (da ein erhöhtes Risiko bei erhöhter Exposition besteht) und sie andererseits couragiert eingreifen und angemessen handeln können, wenn sie einen Vorfall beobachten (Domínguez-Hernández et al., 2018; Fischer et al., 2011; Marín et al., 2019; Rudnicki et al., 2023). Auf diese Weise können potenziell Betroffene besser geschützt und in der Verarbeitung unterstützt werden (z. B. durch soziale Unterstützung). Die Bystander können zudem Selbstwirksamkeit im Umgang mit Hass im Netz erleben (d. h. sie fühlen sich handlungsfähig und in der Lage, im Ernstfall eingreifen zu können), was sie perspektivisch stärkt, falls

sie auch einmal davon betroffen sein sollten. Faktoren, die das Handeln von beobachtenden Personen beeinflussen, lassen sich auf individueller und sozialer Ebene beschreiben (z. B. Domínguez-Hernández et al., 2018; Marín et al., 2019; Rudnicki et al., 2023).

Auf individueller Ebene gelten – vergleichbar zu den Ausübenden – Moral Disengagement, geringe Empathie (v. a. affektiv), geringe Verhaltens-kontrolle und eigene Erfahrung Risikofaktoren für passives Verhalten (d. h. ausbleibende Hilfeleistung bis hin zur Verstärkung der Hassbotschaft). Zudem ist die Wahrscheinlichkeit für den Bystander-Effekt höher, wenn die Gefahreneinschätzung der Situation gering ist, d. h. wenn angenommen wird, das aus der Situation kein relevanter Schaden entsteht, was bei nicht zu beobachtenden Folgen wie psychischer Belastung im Zuge von Hass im Netz wahrscheinlicher ist, da die Folgen nicht direkt beobachtet werden können und für Dritte schwer einzuschätzen sind (Fischer et al., 2011). Daher wird häufig auf eigene Erfahrungen als Referenzrahmen zurückgegriffen. Ent-sprechend sind eigene Viktimisierungserfahrung (v. a. im Fall dort erlebter Unterstützung durch Dritte), stärker ausgeprägte Empathie und ein hohes Handlungswissen Faktoren, die ein aktives Eingreifen und eine Unterstützung der Betroffenen wahrscheinlicher machen können.

Auf sozialer Ebene sind die Beziehungen zu Ausübenden und Betroffenen relevant, da sie z. B. als Peer-Einfluss förderlich für die Unterstützung oder spätere Ausübung von Cyberhass sein können (z. B. bei eigener oder wahrgenommener, sozialer positiver Haltung gegenüber Hass im Netz, etwa im Freundeskreis). Die Beziehungen können allerdings auch präventiv

wirksam werden, wenn z. B. eine enge Freundschaft zu Betroffenen besteht und das aktive Einstehen oder die Gegenrede als Zeichen von Loyalität und Ausdruck der Freundschaft gesehen werden. Da in digitalen Räumen eine Vielzahl von Personen zugleich aktiv ist, kann zudem ein Bystander-Effekt auftreten, wenn es zu Verantwortungsdiffusion kommt, weil die Personen erwarten, dass andere tätig werden und daher passiv bleiben. Im digitalen Raum scheint v. a. die wahrgenommene Anzahl potenziell handelnder Personen dafür entscheidend, eine kritische Masse konnte bisher allerdings nicht bestimmt werden. Durch eine aktive Ansprache von Personen und die Bitte um Unterstützung kann das Auftreten des Bystander-Effekts reduziert werden.

Allgemein hilfreich ist, wenn andere Personen aktiv werden und damit positive Vorbilder sind, um Modelllernen anzuregen (im Sinne digitaler Zivilcourage). Durch die Etablierung und konsistente Umsetzung konkreter Regeln für den Umgang mit Cyberhass kann auf Ebene von Organisationen (z. B. Schulen) dazu beigetragen werden, Bystander-Effekte abzuschwächen und die Bereitschaft zu engagiertem Handeln zu erhöhen.

Zusammenfassung: Perspektive auf Beobachtende

- Vergleichbar mit Mobbing können auch im digitalen Raum Bystander-Effekte festgestellt werden (d. h. anwesende Dritte beobachten Hassphänomene, schreiten aber nicht ein)
- Mögliche Erklärungen sind Einschätzungen geringer Relevanz oder Gefahr bzw. Folgen der Situation, geringe persönliche Verantwortung, fehlendes Wissen oder Ohnmacht in Bezug auf geeignete Handlungs-ansätze
- Auf individueller Ebene sind Moral Disengagement, geringe Empathie und Verhaltenskontrolle sowie Viktimisierungserfahrung Risikofaktoren für Bystander-Effekte
- Auf sozialer Ebene sind Beziehungen zu Ausübenden und Betroffenen, Ausmaß der Moderation (und ggf. Sanktion) der Online-Kommunikation sowie die Verfügbarkeit von (positiven oder negativen) Rollenvorbildern zentrale Einflussfaktoren
- Auf struktureller Ebene kann ein Eingreifen durch klare und verbindliche Regeln in Online-Gemeinschaften sowie einfache Möglichkeiten der Meldung und Kontaktaufnahme gefördert werden
- Wichtig ist, dass Konsequenzen (z. B. Löschung) sichtbar gemacht werden, konsistent erfolgen und als angemessen gelten

Die Rolle von Medium und Umwelt

Als Einflussfaktoren des Mediums und der Umwelt können Aspekte des digitalen Mediums sowie der sozialen und gesellschaftlichen Rahmen-bedingungen gelten. Digitale Medien, die bislang vorrangig beforscht worden sind, sind soziale Medien und Kurznachrichtendienste wie Twitter/X und Facebook/Meta. Plattformen wie Instagram, Snapchat oder TikTok, die insbesondere bei jüngeren Zielgruppen beliebt sind, werden zwar aus theoretischer Perspektive diskutiert, sind empirisch bislang aber weniger beforscht (vgl. Abarna et al., 2022; Chen et al., 2017; Jahan & Oussalah, 2023; Mullah & Zainon, 2021). Gleiches gilt für Nachrichtendienste wie

Telegram, die für empirische Untersuchungen sowie Interventionen weniger leicht zugänglich, Expert*innen zufolge aber hoch relevant für die Entstehung und Praxis von digitalem Hass sind (Castaño-Pulgarín et al., 2021; Mullah & Zainon, 2021; Simon et al., 2022). Ebenso werden Foren wie Reddit, 4chan oder Tumblr diskutiert, die eine Bildung von Interessensgruppen und die Bildung sozialer Identitäten fördern, etwa durch die Entwicklung eigener Rituale, Symbole und Erkennungszeichen und die häufig selbst Nährboden für die Entwicklung von Sub- und Gegenkulturen bieten, da sie andere Nutzer*innen ansprechen als große Plattformen wie Instagram oder Twitter/X (vgl. Castaño-Pulgarín et al., 2021; Costello & Hawdon, 2018; Rieger et al., 2021; Tomczyk et al., 2022).

Die Kommunikation in diesen Interessensgruppen kann die Entwicklung und Aufrechterhaltung von Hass im Netz ermöglichen, ohne dadurch z. B. strafrechtlich relevant zu werden, da etwa Ausdrucksweisen oder Grußformeln häufig nicht mehr rein textbasiert und soweit entfremdet sind (etwa durch Einbindung von Symbolen und Bildinhalten), dass sie weder durch Algorithmen noch durch manuelle Suche eindeutig klassifizierbar sind; vergleichbare Phänomene sind etwa aus der Radikalisierung in digitalen Räumen und Untersuchungen von Online-Extremismus bekannt (Nouh et al., 2019; Sponholz, 2021). Da der Zugang zu diesen Gruppen häufig restriktiv ist und sie eigene Kommunikationsformen entwickeln, die z. B. über spezifische Memes, Sprichworte oder Zeichenkombinationen arbeiten, sind sie für algorithmenbasierte Ansätze schwer zweifelsfrei identifizierbar und zudem für direkte Ansprache (z. B. als Gegenrede) nur bedingt zugänglich.

Vielversprechende Ansätze, etwa aus der Kommunikationswissenschaft, setzen daher stärker auf community-orientierte Ansätze, die stärker partizipativ mit spezifischen Gruppen (z. B. in Subreddits) arbeiten und auf diese Weise Zugang und Verbundenheit schaffen (Hintz & Betts, 2022). Dies erfordert einen gewissen Ressourceneinsatz sowie eine Vertrautheit mit den Plattformen, den Gruppen und ihren Regeln – und damit eine entsprechende Expertise. Zugleich kann an dieser Stelle aber auch Forschung zum Social Influencer-Marketing Berücksichtigung finden, die dabei helfen kann, Botschaften so zu gestalten, dass sie für Nutzende der Plattformen besonders ansprechend, relevant und anregend sind und entsprechend auch von Influencer*innen ernstgenommen und in ihrem Wirkungskreis verteilt werden (vgl. Vrontis et al., 2021 für ein Rahmenmodell zum Social Influencer Marketing). Dieser Zugang ist vielversprechend, in der Präventionsarbeit bisher noch nicht sehr weit verbreitet.

Die zunehmende Verschränkung der Kommunikation durch die Vernetzung verschiedener Plattformen, die Kombination von Online- und Offline-Kommunikation und die Verschlüsselung und Dezentralisierung digitaler Kommunikation ist eine zusätzliche Herausforderung für die Analyse, Prävention und Intervention bei Hass im Netz. Den meisten sozialen Medien, die in diesem Kontext relevant sind, ist gemein, dass sie eine pseudonymisierte und verschlüsselte Kommunikation ermöglichen, die die Identifikation einer Person erschweren (Humphreys, 2018). Durch Anwendungen wie Virtual Private Networks werden zudem weitere Möglichkeiten der Verschleierung geboten, auch wenn hier noch Wissenslücken in der Bevölkerung bestehen (Dutkowska-Zuk et al., 2022). Die wahrgenommene

Anonymität und die geringere soziale Kontrolle (die auch, technisch betrachtet, aufgrund der Vielzahl und Geschwindigkeit an Interaktionen schwer umzusetzen ist), kann zu sozialer Disinhibition im digitalen Raum beitragen (vgl. Online-Disinhibition), sodass Informationen geteilt und Verhaltensweisen gezeigt werden, die außerhalb der digitalen Räume als deviant und sozial unerwünscht gelten können (Suler, 2004; Wachs et al., 2019).

Da sozial unerwünschtes Verhalten in Online-Welten weniger geächtet ist und seltener sanktioniert wird, ist es gerade für Personen mit antisozialen Tendenzen oder geringer Impulskontrolle in diesen Kontexten attraktiver. Die Omnipräsenz sozialer Medien und der Berichterstattung über Hass im Netz stellt eine weitere Herausforderung dar – einerseits soll durch die verstärkte Aufmerksamkeit in der Bevölkerung ein Bewusstsein geschaffen und aufgeklärt werden, was Personen handlungsfähiger machen soll, andererseits wird durch die zunehmende Exposition das Phänomen Hass im Netz normalisiert, sodass es als deskriptive soziale Norm wahrgenommen werden kann, was die Akzeptanz und die Verbreitung steigern kann (Bilewicz & Soral, 2020; Soral et al., 2018).

Ein gutes Verhältnis aus informierender Berichterstattung, Aufklärung und Ansätzen der Problematisierung von Hass im Netz, um zur Denormalisierung des Phänomens beizutragen sollte daher das Ziel sein. Dies ist aber schwerlich in konkrete Empfehlungen übersetzbar, da auch hier weitere Forschung zum Einfluss der Berichterstattung und Rezeption auf psychosoziale Prozesse im Bereich Cyberhass fehlt (Bilewicz & Soral, 2020; Blaya, 2019).

Die Kategorisierung von Personen, die für den gruppen- oder merkmalsbezogenen Ausdruck von Hass relevant sind, wird durch die sozialen Medien erleichtert, da z. B. durch die Aufforderung zur Erstellung von (öffentlich sichtbaren) Profilen mit Präferenzen, Angaben von Interessen, Identitätsmerkmalen (z. B. sexuelle Orientierung) usw. direkt dazu eingeladen wird. Einerseits sollen diese Angaben die Bildung von positiv besetzten Gruppen (z. B. Musikvorlieben, Freizeitaktivitäten) und die Interaktion mit anderen Personen mit ähnlichen Interessen algorithmenbasiert erleichtern, indem besonders interessante Inhalte gefiltert werden, andererseits fördern sie merkmalsbezogene – und damit gruppenbezogene – Bewertungen und Handlungen, die sich z. B. in Hassrede niederschlagen können (Baer, 2019; Mozafari et al., 2020), wenn etwa Filterfunktionen genutzt werden, um Personen zu identifizieren, die diffamiert werden sollen.

Zudem kann die Selektion durch Algorithmen auch dazu beitragen, abwertende oder demokratiefeindliche Inhalte zu verstärken und zu selegieren, wenn sie eine entsprechende Reichweite erzielen und sich statistisch mit Interessen der Person verbinden lassen. Hier sei auf die Debatte zu Filterblasen und Echokammern und deren Implikationen als soziotechnische Mechanismen der politischen Meinungsbildung, Polarisierung und Radikalisierung in sozialen Medien verwiesen (z. B. Bruns, 2019; Figà Talamanca & Arfini, 2022; Interian et al., 2023).

Insgesamt ist viel über Auftreten und Verbreitung plattformspezifischer Formen von Cyberhass bekannt (z. B. in Twitter/X), es existiert allerdings wenig vergleichende Forschung, die Entstehung, Verlauf und Folgen von

Hass im Netz über verschiedene Plattformen, Altersgruppen und Themenfelder hinweg betrachtet: In einer Übersichtsarbeit zum Thema wird der einseitige Fokus auf Plattformen wie Twitter/X sowie das Fehlen der Reflexion der Überschneidung von Hass im Netz und (strukturellem) Rassismus bemängelt (Matamoros-Fernández & Farkas, 2021). Die Autor*innen zeigen auf, dass die meisten Arbeiten in diesem Bereich aus westlich-industrialisierten Ländern stammen (v. a. USA) und zumeist von Weißen durchgeführt werden. Daher sind diverse Perspektiven (z. B. BIPoC[5], LGBTQIA+[6]) zumeist unterrepräsentiert, was zu einer (unbewussten) Verzerrung in der Auswahl, Analyse und Interpretation der Ergebnisse beitragen kann. Zudem findet in der Analyse selten eine Verbindung human- und sozialwissenschaftlicher mit technologiefokussierter Forschung statt, sodass z. B. automatisierte, algorithmenbasierte Auswertungen sozialer Medien nicht ausreichend an Erkenntnisse dieser Forschung angeknüpft und durch sie informiert sind, was ebenfalls zu einer Verzerrung der Ergebnisse beitragen kann.

[5] B(I)PoC ist ein Begriff, der sich auf Schwarze, Indigene und People of Color bezieht und auch als Selbstbezeichnung verwendet wird. Er schließt alle Personen ein, die sich vorrangig als nicht-weiß verstehen und aufgrund dieses Unterschieds Diskriminierung ausgesetzt sind.

[6] Das Akronym LGBTQIA+ steht für die englischen Worte: lesbian, gay, bisexual, transgender/transsexual, queer/questioning, intersex, asexual (übersetzt: lesbisch, schwul, bisexuell, transgender/transsexuell, queer/fragend, intersexuell, asexuell). Das + steht als Platzhalter für weitere Geschlechtsidentitäten.

Zusammenfassung: Medien

- Soziale Medien begünstigen Online-Disinhibition und Deindividuation, die in Zusammenhang mit Cyberhass und Radikalisierung stehen
- Viele Plattformen (z. B. Telegram) sind bislang unterforscht, der Schwerpunkt liegt auf Twitter/X (und z. T. Facebook), sodass viele Gruppen nicht umfassend repräsentiert werden
- Automatisierte Datenanalyse steht vor der Herausforderung der Auswahl unverzerrter und repräsentativer Trainingsdatensätze für maschinelles Lernen sowie der Dynamik von Kommunikationsformen und Soziolekt
- Mediale Berichterstattung kann Hass im Netz normalisieren, daher ist auf ausgewogene Berichte zu achten und Medienschaffende sind ebenfalls als wichtige Akteure im Feld zu berücksichtigen
- Die Forschung sollte Rezeptionseffekte stärker in den Blick nehmen und partizipativ mit Minoritäten arbeiten (z. B. BIPoC, LGBTQIA+)

Ein weiterer Aspekt des Umfelds betrifft die Lebenswelt der beteiligten Personen. So zeigt sich, dass eine geringe soziale Integration und Unterstützung im persönlichen Umfeld sowie positive Haltungen gegenüber Hass im Netz sich förderlich auf die eigene Umsetzung von Hass im Netz auswirken (Chen et al., 2017; Evangelio et al., 2022; Guo, 2016; Henares-Montiel et al., 2022; Li et al., 2021; Lo Cricchio et al., 2021; Petras & Petermann, 2019; Rudnicki et al., 2023; Zych et al., 2019).

Im Kinder- und Jugendbereich sind zusätzliche Risikofaktoren eine negative Eltern-Kind-Beziehung, autoritäre Erziehungsstile und ein restriktiver Umgang im Vergleich zu einem instruktiven Umgang mit Medien (d. h. das Kontrollieren und Verbieten der Mediennutzung gegenüber einer Erläuterung der Gefahren und einer begleiteten Heranführung an die Nutzung) (Domínguez-Hernández et al., 2018; Wachs et al., 2021). In der Zusammen-

schau verweisen diese Befunde damit auf ihre jeweiligen Forschungstraditionen (vgl. Sponholz, 2020 für eine begriffliche Einordnung von "Hate Speech") sowie eine breite Auswahl möglicher Ansatzpunkte für die Prävention und Intervention. Um die Ansatzpunkte weiter einordnen und relevante Mechanismen nachvollziehen zu können werden im Folgenden wesentliche theoretische Zugänge und deren Bedeutung für die Prävention erläutert.

Theoretische Zugänge zur Entstehung und Prävention von Hass im Netz

Als kriminologische Theorien zur Erklärung von Hass im Netz bzw. Cyberhass werden häufig die **Routine Activity Theory** (Cohen & Felson, 1979; Leukfeldt & Yar, 2016; Li et al., 2021) und das **General Aggression Model** (Anderson & Bushman, 2018; Bushman & Anderson, 2002) herangezogen. Die Routine Activity Theory beschreibt als kriminologische Theorie das Zusammenkommen von Personen, die als Täter*in motiviert sind, die als Opfer geeignet erscheinen, einer passenden Situation für die Tat und das Fehlen von geeigneter Begleitung oder Kontrolle, um das Aufeinandertreffen zu entschärfen. Motivierende Faktoren für die Tat können z. B. Vorurteilsstrukturen und ideologische Überzeugungen sein, Merkmale eines potenziellen Opfers z. B. Anzeichen von Vulnerabilität wie junges Alter, physische Schwäche, und als Begleitung kommen z. B. Eltern oder Moderator*innen in Online-Foren in Betracht. Die günstige Gelegenheit bietet sich im Onlinekontext durch schnelle, niedrigschwellige, konsequenzarme und anonyme Form der Kommunikation. Durch eine Bearbeitung zur Tat motivierender Faktoren, die Etablierung geeigneter Begleitstrukturen, die

Stärkung des Beschwerdemanagements und Beschwerdemonitorings zur Veränderung der Rahmenbedingungen einer potenziellen Tat usw. soll daher Hass im Netz vorgegriffen werden.

Das General Aggression Model nimmt eine stärker situationsorientierte Perspektive ein und beschreibt Inputfaktoren (z. B. Eigenschaften von Personen wie Alter, Geschlecht oder Persönlichkeit), situative Faktoren (z. B. wahrgenommene Unterstützung), Prozessfaktoren (z. B. Erregung, Informationsverarbeitung, Impulsivität) und Ergebnisfaktoren (z. B. psychische Belastung, Verarbeitung) in Zusammenhang mit Aggression und Tätlichkeit. Das Model ist bereits auf spezifische Felder wie Cyberbullying übertragen worden (z. B. Kokkinos & Antoniadou, 2019; Kowalski et al., 2014) und bietet durch den prozesshaften Charakter vor allem Anregungen für situationsbezogene Intervention, z. B. zur Emotionsregulation, zur gewaltfreien Kommunikation oder Konfliktbewältigung. Seine Komplexität erschwert allerdings auch die empirische Prüfung und Anwendung des Modells in der Gesamtheit, sodass zumeist einzelne Pfade geprüft werden, die sich in ähnlicher Form auch in vielen anderen Ansätzen wiederfinden.

Dazu gehören z. B. Ansätze, die auf sozial- und motivationspsychologische Theorien rekurrieren, wie die **Theory of Reasoned Action, Theory of Planned Behavior** (Ajzen, 2012; Jung, 2023), die **Social Cognitive Theory** (Bandura, 2001; Chen et al., 2017) oder die **Theorie zur Vorurteilsbildung bzw. zum Intergruppenkontakt** (Imperato et al., 2021; Pettigrew & Tropp, 2006; Soral et al., 2022) und die **Social Identity Theory** (Douglas et al.,

2005; Tajfel & Turner, 1986). Dabei werden sozialpsychologische Erklärungen zur Entstehung von Gruppenidentitäten und Gruppenprozessen herangezogen, die eine Bildung von Vorurteilsstrukturen infolge von Kategorisierungsprozessen postulieren, bei denen negative Zuschreibungen z. B. der Abgrenzung oder Abwertung anderer Gruppen gegenüber der eigenen Gruppe dienen. Dies kann durch ablehnende Worte oder Handlungen auch symbolisch ausgedrückt werden.

Diese Erklärungen werden durch Handlungstheorien motivationspsychologisch durch die Verknüpfung von Situation, Motivation und Ziel ergänzt, die genauer in den Blick nehmen, welche individuellen und sozialen Motive in einer Situation auftreten, welche Handlung diese Motive erfüllt und wie sie damit zur Erreichung eines bestimmten Ziels beiträgt (z. B. Emotionsregulation, Popularitätsgewinn). Aus diesen Traditionen heraus lassen sich viele Hinweise zur Arbeit an Ziel- und Motivationsklärung, Erarbeitung alternativer Strategien zur Zielerreichung, Schaffung positiver Gruppenidentitäten und zum Aufbau von Kontaktinterventionen mit den betroffenen Fremdgruppen ableiten, die entsprechend in Individual- oder Gruppeninterventionen Eingang finden.

Speziell für den Bereich beobachtender Personen (*Bystander*) wurde ein **Bystander Intervention Model** erarbeitet, das Personen befähigen soll, eine aktive, helfende Rolle einzunehmen und den Bystander-Effekt zu reduzieren (Bennett et al., 2014; Fischer et al., 2011; Latané & Darley, 1968). Es führt zentrale Variablen verschiedener Theorien und Ansätze zusammen und leitet daraus Handlungsimpulse ab: Um erfolgreich intervenieren können, muss ein

Vorfall zunächst wahrgenommen werden (*Notice*), als relevant erachtet werden (*Interpret*), eigene Verantwortung für das Handeln muss akzeptiert werden (*Accept*), hilfreiche Handlungsansätze müssen bekannt sein (*Know*) und Handlungen müssen dann auch umgesetzt werden (*Act*). Maßnahmen, die auf diesem Modell beruhen, versuchen daher, bei den Teilnehmenden ein Bewusstsein für Cyberhass zu schaffen, Wissen über relevante Anzeichen und Ansatzpunkte für Unterstützung aufzuzeigen, verschiedene Strategien zum Umgang mit Cyberhass zu erarbeiten und zu erproben, sodass eine Umsetzung im Ernstfall erleichtert werden sollen (vgl. Bennett et al., 2014; Fischer et al., 2011; Lan et al., 2022; Latané & Darley, 1968; Obermaier, 2022)

Aus kommunikations- und rechtswissenschaftlicher Perspektive sind die **Critical Discourse Analysis** (al-Utbi, 2019; Kress, 1990) sowie die **Critical Race Theory** (Bliuc et al., 2018; Delgado & Stefancic, 1998) wichtige und häufig genutzte Bezugspunkte für das Verständnis und die Prävention von Hass im Netz. Die Critical Discourse Analysis erläutert, wie Hass im Netz Ausdrucksform alltäglicher, erlebter Ereignisse sein kann und durch Handlungen mentale Hassmodelle und Erlebnisse reproduziert. Die Critical Race Theory richtet ihren Blick auf gesellschaftliche Machtverhältnisse und Ungleichheit und legt nahe, dass Hass im Netz, z. B. als Hassrede, Form symbolischer Gewalt ist, die tradierte Verhältnisse reproduziert, denn zu den Betroffenen zählen vorrangig (historisch) marginalisierte Gruppen. Diese Ansätze bieten z. B. Anlass, sich mit den Erfahrungen und Bewältigungs-mechanismen von Hass im Alltag auseinanderzusetzen, dazu zählt im Besonderen auch die Alltagssprache und -kommunikation, um deren Transfer

in den Online-Kontext besser verstehen und ggf. frühzeitig intervenieren zu können. Zudem laden sie zu einer kritischen Reflexion der eigenen Positionalität und der gesellschaftlichen, politischen und sozialen Strukturen ein, in denen Hass im Netz stattfindet.

Schließlich liefern auch medienwissenschaftliche und medienpädagogische Ansätze wichtige Hinweise zur Rolle von Medien im Umgang mit Cyberhass, so zeigen etwa Studien zu **Media Effects** (Chen et al., 2017; Suler, 2004; Valkenburg et al., 2016), dass Onlinemedien durch fehlende Regulation und Begleitung sowie Möglichkeiten der Anonymität Disinhibition begünstigen können. Dies ist wiederum mit erhöhter Gewaltbereitschaft und -äußerung im digitalen Raum und damit auch Hass im Netz assoziiert.

Eine wesentliche Grundlage für erfolgreiche Prävention ist demnach eine Stärkung der Medienkompetenz, d. h. des Wissens über Medien, die kritische Reflexion der eigenen Mediennutzung und der in Medien ablaufenden Prozesse sowie der kreativen Nutzung von Medien (vgl. Baacke, 1996; Jeong et al., 2012). Hinzu kommt die ethische Mediennutzung im Sinne der demokratieförderlichen und dem Normsystem entsprechenden Erwartungsrahmen, z. B. in Form von Digital Citizenship (Choi, 2016). Basierend auf diesen theoretischen Ansätzen sowie den empirischen Befunden lässt sich eine Vielzahl an Interventionsmöglichkeiten beschreiben, die die Verhältnisse und das Verhalten in den Blick nehmen.

Zusammenfassung: Theoretische Zugänge

- Kriminologische Theorien beschreiben v. a. personenbezogene und situative Faktoren, die zu Hass im Netz führen können
- Handlungstheoretische und sozialpsychologische Theorien fokussieren den Prozess einer Handlungsentscheidung und beschreiben, welche sozialen Einflüsse darauf einwirken
- Soziologische Theorien stellen die Kontextualität von Hass im Netz heraus und diskutieren die Relevanz weiterer Trends (z. B. Digitalisierung) in diesem Kontext
- Kommunikationswissenschaftliche Theorien diskutieren die Bedeutung tradierter Deutungsmuster für die v. a. verbale Gestaltung von Hass im Netz und die Reproduktion von Machtverhältnissen und Ungleichheit
- Medienwissenschaftliche Theorien beschreiben die Interaktion aus Medium, Gegenstand und Personen und die Rolle von Medienbildung

Verhältnis- und verhaltensbasierte Ansätze der Prävention und Intervention

Verhältnisbasierte Ansätze adressieren die Rahmenbedingungen von Verhalten und betreffen im Kontext von Hass im Netz etwa die Gesetzgebung, die Ausdrucksmöglichkeiten im digitalen Raum oder die Regeln für die Nutzung digitaler Medien (Mair & Mair, 2003). Das seit 2017 bestehende Netzwerkdurchsetzungsgesetz und der Digital Services Act zielen bspw. auf eine Stärkung der Rechte von Nutzer*innen sozialer Medien ab durch transparentes, verständliches Beschwerdemanagement (inkl. Reports), Auskunfts- und Informationspflichten von Anbieter*innen und Sanktionsmaßnahmen wie Bußgelder. Dies sollte seit 2021 durch eine Meldepflicht schwerer Straftaten auf Grundlage des Gesetzes zur Bekämpfung des Rechtsextremismus und der Hasskriminalität ergänzt werden, die die Strafverfolgung

erleichtern und damit auch die Abschreckung von digitaler Hasskriminalität unterstützen soll.

In der Praxis zeigen sich einige Herausforderungen der Umsetzung, so etwa die Entwicklung und Erprobung unverzerrter und zuverlässiger Algorithmen, die eine eindeutige Identifikation von Hassbotschaften erlauben, die Bekanntheit und Nutzerfreundlichkeit von Beschwerdemöglichkeiten und Auskunft oder die Definition relevanter Korpora für die Identifikation von Hassrede. Obgleich umfangreiche Lexika wie *HurtLex* (Poletto et al., 2021) für die (teil-)automatisierte Identifikation entwickelt worden sind, basieren sie auf selektiv gewählten Daten (v. a. Twitter, Facebook, Reddit), sind schnell veraltet, da sie die Codierung und Neologismen der Online-Diskurse relevanter Gruppen nicht angemessen reflektieren können (s. o.), die korrekte Verarbeitung nichttextlicher Inhalte herausfordernd ist (insbes. videobasierte Inhalte oder Emojis) und die Unterschiede kultureller und nonverbaler Aspekte der Sprache nicht verlässlich reproduziert werden können, wie Sarkasmus oder Idiome (Abarna et al., 2022; Khurana et al., 2023; Sarsam et al., 2020). Eine (teil-)automatisierte Datenanalyse ist überdies auf unverzerrte, repräsentative und verlässliche Trainingsdatensätze angewiesen, die aufgrund der Dynamik der Diskurse, der Vielfalt der Themen und Ausdrucksformen schwer zu gewährleisten ist. Neben der Weiterentwicklung von Analysemethoden kann eine systematische, interdisziplinäre, an Replikation und Metasynthese orientierte Forschung zur Qualitätssicherung beitragen, die i. S. offener Forschung (*Open Science*) Datensätze, Software und Analysemethoden teilt, Analysen repliziert und kritisch reflektiert (vgl. Garland et al., 2022).

Eine weitere Herausforderung ist die Umsetzung der gesetzlich definierten Verpflichtungen durch Anbieter*innen, da zu kontrollieren ist, wie z. B. Beschwerdemanagement organisiert, koordiniert und kommuniziert werden, wie Beiträge gefiltert und gelöscht werden, wie die Einbindung künstlicher Intelligenz in den Prozess gestaltet wird (z. B. über Chatbots). Bei Nichteinhaltung müssen auszusprechende Sanktionen auch nicht unmittelbar zu Änderungen in der Praxis führen, da deren Durchsetzung ebenfalls Zeit benötigt. Eine konsequente Durchsetzung dieser Aspekte sowie die Dokumentation und Nachverfolgung kann den Prozess potenziell verbessern, erfordert aber, etwa im Kontext der Strafverfolgung, eine hohe Fachkompetenz, fortlaufende Kommunikation und entsprechend enorme personelle und finanzielle Ressourcen.

Ein weiterer verhältnispräventiver Ansatzpunkt ist die Bemühung, vorurteils- und hassfreie digitale Räume zu schaffen, indem z. B. Foren und Gruppen mit Selbstverpflichtung, Kodizes und Moderation arbeiten, die eine hassfreie Umgebung sicherstellen und für das Thema sensibilisieren soll. Wenngleich dieser Ansatz der Selbstverwaltung die Verantwortung von Anbieter*innen zu den Nutzer*innen verschiebt und einen Mehraufwand für diese bedeutet (z. B. Zeit, Qualifikation, Dokumentation), so ist es besonders für Personen in vulnerablen Situationen (z. B. Kinder, Menschen mit geringen Sprachkenntnissen) eine gute Möglichkeit, digitale Kommunikation zu erleben. Das enorme zivilgesellschaftliche Engagement in diesem Bereich ist Zeichen hoher Akzeptanz und Bereitschaft, dies anzugehen. Gleichwohl ist fraglich, inwiefern ein solches Vorgehen nachhaltig skalierbar ist und welche Kompetenzen für die Kommunikation außerhalb dieser Räume erworben

werden müssen oder können (etwa in der Verschränkung von Online- und Offline-Hass oder im Umgang mit unterschiedlichen Formen und Kontexten von Hass im Netz). Zudem ist unklar, wie der Mehraufwand langfristig getragen werden soll, bislang erfolgt dies v. a. durch ehrenamtliches, zivilgesellschaftliches Engagement.

Neben diesen verhältnispräventiven Ansätzen, die den Rahmen für Online-Kommunikation schaffen, zielen die meisten Ansätze zu diesem Themenkomplex auf Verhaltensprävention ab, um entweder beobachtende, ausführende oder betroffene Personen direkt anzusprechen oder relevante Personen im sozialen Umfeld zu adressieren (z. B. Lehrkräfte, Peers, Eltern). Wesentliche Ziele sind, Medienkompetenz, digitale Kompetenzen und Partizipation zu stärken (dazu zählt auch Wissen zu Cybersicherheit), alternative Strategien, z. B. der Emotionsregulation, zu erarbeiten und Kompetenzen im Umgang mit Hass im Netz (z. B. Beschwerdemanagement) zu fördern und zum Austausch anzuregen (z. B. durch Fortbildungen für Pädagog*innen oder Fachkräfte und Gesprächsgruppen für Schüler*innen).

Im Zuge der Erstellung dieses Gutachtens wurden sechzehn Angebote identifiziert und analysiert (HateLess, Helden statt Trolle, Fairplayer.Manual, Medienhelden, Surf-Fair, BITTE WAS?!, Zivile Helden/PräDiSiKo, Law4School, Cybermobbing Prävention e. V., Klicksafe.de, Juuuport e. V., ChatScouts, No Hate Speech Movement, Love-Storm, Courage im Netz, Firewall – Hass im Netz begegnen), für die mindestens theoretisch gut begründete Wirksamkeitsannahmen bestehen, manualisierte oder zumindest

strukturierte Vorgehensweisen dokumentiert sind und z. T. Evaluationsstudien vorliegen. Da das Ausmaß der Strategien, Dokumentation und Evaluation für diese Angebote jedoch stark variiert und die Angebote vielfach kontextabhängig sind (z. B. wurden sie in oder für bestimmte(n) Schularten oder Gruppen entwickelt), lassen sich schwerlich allgemeine Empfehlungen für einzelne Angebote aussprechen, ohne dabei Einschränkungen, etwa hinsichtlich der Evidenzgrade, der Wirksamkeit, der Übertragbarkeit (z. B. auf verschiedene Bundesländer, Schulkontexte, Zielgruppen) zu bedenken und wichtige Schritte der Adaptation und Implementation von Maßnahmen zu betonen. Zudem laufen derzeit noch einige Evaluationsstudien (z. B. zur Wirksamkeit von Law4School), die bei der Auswahl und Gestaltung ebenfalls berücksichtigt werden sollten. Daher wäre ein systematisches Mapping der Best Practices, Evidenzgrade und relevanter Hinweise zur Adaptation, Implementation und Evaluation dieser Angebote ein wichtiger nächster Schritt. Im Folgenden werden daher zusammenfassend zentrale Ansatzpunkte und Ergebnisse berichtet, sie sich in der nationalen und internationalen Forschung übergreifend und mehrheitlich als relevant herausgestellt haben, ohne dabei spezifische Programme als besonders (wenig) empfehlenswert herauszustellen.

Als wirksame Ansätze (d. h. Programme, für die mindestens eine randomisierte kontrollierte Studie oder eine rigoros umgesetzte quasi-experimentelle Studie zur Wirksamkeitsprüfung vorliegt) haben sich dabei settingbasierte, strukturierte Programme (z. B. im Schulkontext, im Vergleich zu Freizeitangeboten) von längerer Dauer (mind. 8 Stunden) erwiesen, die einen Methodenmix nutzen (Präsenz + digitale Anwendung wie virtuelle Realität,

webbasierte Anwendungen, Nachrichtendienste wie WhatsApp) (z. B. Doty et al., 2021; Mishna et al., 2009; Polanin et al., 2022). Inhaltlich konnten Wirksamkeitsnachweise für Bildungsangebote zu Cyberhass, die Stärkung von affektiver Empathie, sozio-emotionalen Kompetenzen, Kommunikation und Training zur Teilhabe an digitalen Medien (Digital Citizenship, dazu zählen u. a. Aspekte der Netiquette, Partizipations- und Gefahrenpotenziale digitaler Medien) (Choi, 2016) erbracht werden.

Digital Citizenship beschreibt die Wahrnehmung einer gesellschaftlichen Rolle eines Individuums durch digitale Technologien (Hintz et al., 2017) und erstreckt sich entlang des Kontinuums digitaler Partizipation. So umfasst eine Bildung i. S. des Digital Citizenship etwa Informationen und Aufklärung zu relevanten Phänomenen und Prozessen (wie Filterblasen oder Echokammern) und Kompetenzen zur Nutzung digitaler Medien, damit bestehen Gemeinsamkeiten zur Medienkompetenz. Der Fokus liegt hier allerdings auf gesellschaftlicher Teilhabe und damit verbundenem Aktivismus (vgl. George & Leidner, 2019). So können Personen bereits durch das Betrachten von (politischen) Inhalten und niedrigschwellige Interaktion (z. B. Likes) partizipieren, darüber hinaus aber auch durch Crowdfunding oder digitale Petitionen und Meinungsaustausch sowie durch datenbezogene Aktivitäten (z. B. digitale Selbstverteidigung, vgl. https://digitalcourage.de/) wirksam werden. Da Digital Citizenship mit einem Verständnis von Grundrechten verbunden ist und in Deutschland z. B. die Verbundenheit mit der freiheitlichen demokratischen Grundordnung impliziert, besitzt sie Bezüge zur Demokratie- und Medienbildung. Das Konzept ist vergleichsweise jung, weshalb es bislang weniger Forschung zu Auswirkungen von Digital

Citizenship auf Hass im Netz gibt, aber die angesprochenen Kompetenzen finden sich bereits in einigen Ansätzen wieder (z. B. Angeboten der Medienpädagogik oder Demokratieförderung).

Im Bereich der Elternarbeit haben sich Aufklärung über Cyberhass und assoziierte Phänomene, Stärkung der Erziehungskompetenz hinsichtlich instruktiver gegenüber restriktiver Medienerziehung und Stärkung der Eltern-Kind-Bindung als förderlich erwiesen (Hutson et al., 2018; Mishna et al., 2009; Mula Falcón & Cruz González, 2023; Wachs et al., 2021). Allerdings sind diese Ansätze häufig isoliert betrachtet worden und selten Teil komplexer Interventionen, die neben elterlichen Zielgruppen weitere Zielgruppen und Kontexte in den Blick genommen haben, z. B. die Wechselwirkungen von Medienerziehung durch kombinierte Interventionen bei Eltern und Kindern, den Einfluss auf Mediennutzung in der Familie, den Austausch und die Vernetzung mit anderen Lebensbereichen (z. B. Jugendarbeit, Schule, Freizeit). Die Effektstärken für verhaltensbasierte Interventionen liegen überwiegend im positiven sowie kleinen bis mittleren Bereich. (z. B. Doty et al., 2021; Mishna et al., 2009; Polanin et al., 2022; Soral et al., 2018). Somit kann für viele Ansätze festgestellt werden, dass sie Einstellungen und zumeist selbstberichtetes Verhalten im Bereich Hass im Netz beeinflussen können.

Weniger Evidenz liegt bisher für rein digitale Interventionen oder heim- und alltagsbasierte Interventionen vor, die z. B. eigenständig per App bearbeitet werden können und die dem Alltag angepasst sind. Auffällig ist auch, dass viele Interventionen Individuen ansprechen und dabei soziale Gruppen oder

relevante Personen aus dem sozialen Umfeld (z. B. Eltern, Freundeskreis) nicht oder nur sporadisch und häufig passiv eingebunden werden (z. B. durch die Weitergabe von Informationen, Informationsveranstaltungen) (Lan et al., 2022; Mula Falcón & Cruz González, 2023; Wachs et al., 2023). Dies kann eine Erklärung für die überwiegend kleinen Effektstärken sein, da viele kontextbezogene und interpersonale Faktoren außer Acht gelassen werden, und verweist auf Entwicklungspotenzial in der Gestaltung präventiver Ansätze: Durch eine stärker an der Gemeinschaft orientierte, partizipative Gestaltung kann eine größere Bereitschaft des Mitwirkens relevanter Akteure (z. B. Jugendarbeit, Eltern, Lehrkräfte, Peers, Polizei) erreicht werden, um frühzeitig Personen und Gruppen identifizieren und ansprechen zu können, die mit dem Thema in Berührung kommen und ihnen gezielter passende Unterstützung zukommen lassen zu können. Für relevante Lebenswelten können das jeweils unterschiedliche Akteure sein, so z. B. Influencer*innen im Bereich sozialer Medien (Vrontis et al., 2021), Familie und Angehörige im sozialen Nahraum, Peers im Schulkontext, digitale Freiwillige und Community Managers in spezifischen Online-(Interessens-)Gruppen und Personen in Freizeiteinrichtungen (z. B. Sportvereine, Jugendarbeit).

Zentrale Komponenten der erfolgreichen Interventionen sind Kompetenz-aufbau durch Wissensvermittlung, Bearbeitung von Lern-Lehr-Material (z. B. Texte zu Fallbeispielen lesen und bearbeiten), Psychoedukation (z. B. Aufklärung über Grundlagen der Stressregulation, psychosoziale Folgen von Hass im Netz) und Training (z. B. Rollenspiele). Seltener eingesetzt werden Komponenten zur Veränderung des systemischen Umgangs mit dem Thema (z. B. Regeln oder Interventionen zur Beeinflussung von Schul- oder

Betriebsklima) sowie das gezielte Erarbeiten und Trainieren spezifischer Reaktionen, z. B. das aktive Gestalten von Gegenrede. Interventionen, die sich spezifisch mit der Gestaltung und Verbreitung von Hass im Netz auseinandersetzen, um diese Lücke zu schließen, sind bislang kaum evaluiert – so konnten in einer kürzlich erschienenen Übersichtsarbeit (Windisch et al., 2022) nur zwei Studien in diesem Feld identifiziert werden, davon eine aus Deutschland, die beide keine statistisch bedeutsamen Effekte berichtet haben. Diese Befunde stimmen auch mit der Kritik von Blaya (2019) überein, wonach zwar zahlreiche Aktivitäten und Maßnahmen zur Gestaltung von Gegenrede existieren, für diese aber keine belastbaren Wirksamkeitsnachweise existieren.

Daraus kann nicht abgeleitet werden, dass Interventionen zu Gegenrede keine Wirkung besitzen, aber der Beitrag verweist auf Schwierigkeiten im Feld. So existieren vielfältige Definitionen von Strategien und Formen der Gegenrede, die sich sehr unterschiedlich auf Hassphänomene auswirken können und die unterschiedlich eingesetzt werden. In einer zum aktuellen Zeitpunkt (Februar 2024) unveröffentlichten Arbeit (d. h. ohne Peer Review zur Qualitätssicherung) berichten Jia und Schumann (2023) beispielsweise, dass Gegenrede, die der Ablenkung auf ein anderes Thema oder der Belehrung der Ausübenden dienen soll, dazu beitragen kann, dass Hassbeiträge eher ignoriert werden. In der Wahrnehmung anderer kann das wiederum zur Normalisierung von Hassrede und der „Nichtreaktion als Reaktion“ beitragen. Eine aggressive Gegenrede, die z. B. Ausübende zurechtweist, kann zum Flaming beitragen und damit selbst die Entstehung von Hassphänomenen begünstigen. Damit hätten solche Interventionen

unerwünschte Nebenwirkungen. Daher ist weitere Forschung notwendig, die verschiedene Arten von Gegenrede genauer betrachtet (z. B. direkte und indirekte Gegenrede, Kommunikation in privaten und öffentlichen Räumen), Determinanten und Prozesse in den Blick nimmt und mit Fokus auf die Prävention prüft, welche Strategien in welchen Situationen für welche Gruppen welche Konsequenzen hervorrufen. Dabei sind insbesondere auch potentiell unerwünschte Konsequenzen mitzudenken.

Eine weitere Herausforderung des gesamten Feldes ist, dass viele Interventionen wissenschaftlich entwickelt und z. T. hinsichtlich ihrer Wirksamkeit untersucht worden sind, in den meisten Fällen aber keine Prüfung der Wirksamkeit unter Alltagsbedingungen stattfand (*Effectiveness*), sodass deren Übertragbarkeit unklar ist (z. B. Blaya, 2019; Polanin et al., 2022; Seemann-Herz et al., 2022). Ebenfalls fehlen Studien zur Implementation und Nachhaltigkeit der Interventionen, die den Wert für die öffentliche Gesundheit überprüfen und den Public Health Impact in den Blick nehmen (vgl. Harden et al., 2018; Kwan et al., 2019). Um eine belastbare Einschätzung von Kosten, Effektivität/Wirksamkeit und Nutzen vornehmen zu können, sind daher weitere Forschungsarbeiten erforderlich. Da zum aktuellen Zeitpunkt für viele Interventionen keine Angaben zu den Kosten verfügbar sind, ist dies eine wichtige Aufgabe für die Zukunft.

Modelle wie das sozioökologische Entwicklungsmodell (Bronfenbrenner, 1977) bieten dafür eine sehr gute Grundlage, um auf Mikroebene (z. B. Individuum, Familie), Mesoebene (z. B. Schule, Nachbarschaft) und Makroebene (z. B. Medien, Gesellschaft) Einflussfaktoren und Prozesse in

Bezug auf Hassphänomene zu charakterisieren und entsprechende Handlungsansätze abzuleiten. Die meisten bisherigen Ansätze befassen sich allerdings mit einzelnen Ebenen, v. a. der Mikroebene (etwa durch den Aufbau individueller Kompetenzen), oder der Makroebene (etwa durch Gesetzesänderungen). Schulbasierte Ansätze, der Definition nach eine mögliche Mesoebene, nehmen vielfach ebenfalls Individuen in den Blick (z. B. Personen in einer Schulklasse) und weniger die Regeln, Strukturen und Kulturen der Schule selbst. Demzufolge sind die Wechselwirkungen dieser Ebenen (z. B. Reaktionen auf Änderungen in Kommunikationsverhalten oder Sanktionsmechanismen) sowie zeitliche Verläufe (z. B. bedeutsame Entwicklungsprozesse im Übergang von Kindheit zur Jugend und die sich damit verändernden Sozialisationskontexte, Einfluss digitaler Innovation und sich verändernder Sprache auf das Verhalten) bislang vernachlässigt (Banyard, 2011; Fischer et al., 2011; Lan et al., 2022).

Im Bereich Cyberbullying – und zum aktuellen Zeitpunkt auch Cyberhass und Cyberaggression – existieren zahlreiche Angebote, v. a. für Schüler*innen, die individuelle Kompetenzen fördern und größtenteils das unmittelbare Umfeld (Peers, z. T. Lehrkräfte) einbinden, sektorenübergreifende Inhalte, die z. B. Lehrkräfte und Eltern zum Thema Strafverfolgung und Sanktionen informieren, mit relevanten Anlauf- und Meldestellen verbinden und in der Moderation von Online-Räumen oder der Nutzung relevanter Plattformen schulen sowie integrative Ansätze zur Zusammenarbeit von kommunalen Akteuren (z. B. Sozialarbeit, Familienhilfe), Eltern, Lehrkräften und Schüler*innen mit Blick auf kommunale Ziele hinsichtlich Cyberaktivitäten sind selten (Lan et al., 2022). Um nachhaltig Veränderungen erzielen zu

können, die gesellschaftlich eine Breitenwirkung entfalten können, ist daher eine Kapazitätsentwicklung zu empfehlen, wie sie etwa in kommunaler Gesundheitsförderung und Kriminalprävention, praktiziert wird, z. B. im Communities that Care-Ansatz (z. B. Kuklinski et al., 2021; Nickel & von dem Knesebeck, 2020; Walter et al., 2023).

Zusammenfassung: Präventions- und Interventionsansätze

- Verhältnisbasierte Ansätze der Prävention adressieren die Rahmenbedingungen von Verhalten (z. B. Gesetze, Richtlinien) und sind in Deutschland weit fortgeschritten (z. B. NetzDG)
- Einige Möglichkeiten (z. B. Anlaufstellen) sind noch nicht hinreichend bekannt und die Umsetzung ist durch fehlende Mittel erschwert (z. B. geeignete Algorithmen, fachlich qualifiziertes Personal)
- Zivilgesellschaftliches Engagement ist vielversprechend in der Gestaltung positiver Kommunikationsräume (z. B. Love-Storm)
- Verhaltensbasierte Ansätze (z. B. Präventionsprogramme) sprechen v. a. individuelle und soziale Prozesse an, die zu weniger Hass im Netz führen sollen, das tatsächliche Auftreten von Hassphänomenen (im Vergleich zum Selbstbericht von Teilnehmenden) wird allerdings bisher selten als Ergebnisparameter betrachtet
- Evidenzbasiert und -informiert sind Ansätze zur Förderung von Empathie, Emotionsregulation, sozialer Kompetenz, Netiquette, Medienkompetenz, Erziehungskompetenzen, die Information mit praktischen Übungen und Reflexion in Gruppen verbinden
- Unzureichend untersucht sind partizipative Ansätze, Studien im Längsschnitt unter Alltagsbedingungen sowie die Adaptation und Implementation evidenzbasierter Intervetionen

Die Situation in Niedersachsen

In Niedersachsen werden Erfahrungen mit Hass im Netz regelmäßig erhoben, so etwa im Niedersachsensurvey, der seit 2013 als regelmäßige, bundesland-weite Befragung von Schüler*innen der 9. Klasse durch das Kriminologische Forschungsinstitut Niedersachsen e. V. durchgeführt wird (https://kfn.de/forschungsprojekte/schuelerbefragungen/). Die Befragung legt im Online-Bereich einen Fokus auf Cyberbullying, -viktimisierung und -kriminalität. In der Zusammenschau bisheriger Studien (z. B. Baier et al., 2019; Bergmann, 2022; Bergmann & Baier, 2018; Bergmann et al., 2018; Krieg et al., 2022) zeigt sich eine große Bandbreite in der Prävalenz, von häufigem Cyberbullying (ca. 1-2 %) bis hin zu seltenem Cyberbullying (bis zu 26 %) sowie von selten (ca. 4 %) bis hin zu häufig erlebter Cyberviktimisierung (bis zu 40 %). Auffällig ist, dass das Verhältnis ungefähr 1:2 ist, sodass womöglich eine höhere Dunkelziffer besteht und Personen, die Cyber-bullying ausüben, dieses z. B. aus Scham seltener berichten. Als Zielgruppen wurden bislang v. a. Schüler*innen der Mittel- und Oberstufe untersucht, die durch Instant Messaging-Dienste Erfahrungen mit Cyberbullying oder Cyberviktimisierung gemacht haben.

Als Risikofaktor für Viktimisierung zeigt sich übergreifend Migrations-hintergrund, für Bullying männliches Geschlecht (hinsichtlich der Häufigkeit, weniger der Intensität oder Form), externalisierendes Verhalten und geringe sozio-emotionale Kompetenzen. Weibliche Personen scheinen eher Cyber-bullying als klassisches Bullying auszuüben. Als Reaktionen auf Cyber-bullying werden v. a. Ohnmacht, Stress und Anspannung und teilweise

depressive Symptomatik sowie sozialer Rückzug berichtet. Personen, die Erfahrungen mit Mobbing außerhalb von digitalen Räumen gemacht haben, als ausübende oder betroffene Person, haben ein erhöhtes Risiko, auch in digitalen Räumen Cybermobbing oder Cyberviktimisierung zu erfahren.

Im Ergebnis erscheint die Situation von Schüler*innen in Niedersachsen vergleichbar mit den o. g. Befunden aus nationalen und internationalen Studien. Allerdings sind die Stichproben häufig selektiv, querschnittlich (sodass keine Kausalaussagen möglich sind) und die Instrumente nicht immer psychometrisch geprüft und validiert. Zudem werden als Erfahrungskontexte häufig schulische Referenzrahmen angenommen und es werden Erfahrungen mit Plattformen wie Snapchat angegeben, die mit bisherigen Studien zu Facebook/Meta und Twitter/X schwer vergleichbar sind. Weitere Lebenswelten wie Freizeit oder Familie sind im Vergleich insgesamt seltener untersucht worden (z. B. elterliche Medienerziehung als Proxy für den familiären Umgang mit Medien und Hass im Netz), sodass eine lebensphasensensible und kontextübergreifende Betrachtung erschwert ist. Es wird deutlich, dass weitere Zielgruppen, z. B. Erwachsene und Menschen, die häufiger von Hass im Netz betroffen sein können (etwa Personen aus Politik oder Strafverfolgung und Behörden, die gezielt Anfeindungen im Netz ausgesetzt sind), bislang eher unterrepräsentiert und weniger rigoros untersucht worden sind. Schließlich sind die Datensätze bisher nicht für die Forschung verfügbar (*Open Data*), sodass eine Reanalyse, Zusammenführung der Daten nicht möglich ist – gerade mit Blick auf die Erstellung von Trainingsdatensätzen für maschinelles Lernen wäre dieses unbedingt zu empfehlen.

Implikationen für die Forschung und Praxis

Aus der Zusammenschau der dargestellten theoretischen Zugänge, empirischen Befunde und infolge einer konsensuellen Validierung (in Form von Expert*inneninterviews) ergeben sich zentrale Ansatzpunkte für die weitere Forschung und Praxis, die auf Ebene der Individuen, Gruppen, Lebenswelten/Settings und Strukturen formuliert werden können.

Mit Blick auf **Individuen**, die Hass im Netz ausüben (wollen), erleben oder beobachten, kann aus der Forschung zu Präventions- und Interventionsansätzen empfohlen werden, sozio-emotionale Kompetenzen zu stärken, um alternative Formen der Emotionsregulation zu etablieren und sich in sozialen Situationen gut behaupten und angemessene Unterstützung einholen oder anbieten zu können. Zudem ist zu empfehlen, an der Stärkung von Medienkompetenz, Digital Citizenship, Selbstwert und Empathie zu arbeiten, etwa durch strukturierte edukative und medienpädagogische sowie psychologische Programme oder im Rahmen individueller Fördermaßnahmen, etwa in Settings von Bildung und Gesundheit. Viele der o. g. Angebote zielen auf die Stärkung dieser Aspekte ab.

In Kontext von **Gruppen** ist zu empfehlen, Gruppennormen in Bezug auf Hass im Netz stärker in den Blick zu nehmen und auf positive Weise wirksam zu werden, z. B. durch die Etablierung positiver sozialer Identitäten und sozialer Normen, die nicht durch Ausgrenzung anderer Personen oder bestimmter Merkmale (z. B. Geschlecht, Migrationshintergrund) definiert sind, sondern andere Gemeinsamkeiten betonen (z. B. gemeinsame Interessen im Bereich Kultur). Dies sollte partizipativ erfolgen, unter Berücksichtigung

und Teilhabe von Personen, die besonders von Hass im Netz betroffen oder für das Ausüben oder Erleben von Hass im Netz vulnerabel sind. Da Hass im Netz vielfach durch frühere Erfahrungen gespeist ist und sich Gruppenstrukturen und -prozesse reproduzieren können, sollten Schnittstellen zwischen digitalen und nicht-digitalen Welten stärker in den Blick genommen werden. Beide Aspekte sind zudem z. B. in der Präventionspraxis zu beachten, da doppelt betroffene Personen auch stärkere Belastungen berichten (z. B. Menesini et al., 2012; Ossa et al., 2023).

Zudem sind die Besonderheiten digitaler Räume (z. B. Anonymität, Disinhibition, soziotechnische Selektion) auch für die Betrachtung von Gruppen zu berücksichtigen, da sowohl die Konfiguration als auch die Kommunikation und der Zusammenhalt von Gruppen einer stärkeren Dynamik unterworfen sind als in Offline-Kontexten. Für die Umsetzung von Programmen sollten zunehmend soziale Gruppen und Interaktionen in den Fokus rücken (im Speziellen auch Online-Gemeinschaften/Communities, die keine Entsprechung in Offline-Kontexten haben), da viele individuumszentrierte Ansätze wie Bystander-Interventionen diese bislang unzureichend berücksichtigen. Dabei kann mit Online-Gemeinschaften oder Interessensgruppen gearbeitet werden, um gezielt soziales Lernen anzuregen und individuelles Lernen zu ergänzen (Lan et al., 2022). Die partizipative Arbeit mit Gruppen, auch in der Analyse und Interpretation von Hass im Netz sowie der Eruierung von Konsequenzen und der Erarbeitung von Gegenstrategien ist dabei zentral, um positiv wirksame und tragfähige soziale Normen zu entwickeln (Lan et al., 2022; Wachs et al., 2023).

Auf Ebene der **Lebenswelten und Settings** kann vor allem für Eltern, Familien und Schulen die Empfehlung abgeleitet werden, Medienbildung und -erziehung zu stärken, sowohl als ganzheitliches Verständnis von Medienkompetenz als auch als Digital Citizenship mit Bezug zur digitalen Partizipation und Selbstbestimmung (s. o.). Ein offener, lernorientierter und instruktiver Zugang erscheint dabei als hilfreich. Die Anregung eines Austausches zum Thema sowie das Schaffen gezielter Räume, um über Hass im Netz zu sprechen und z. B. alternative Strategien zu Hassrede sowie individuelle und kollektive Bewältigungsstrategien zu erarbeiten, bergen ebenso Potenzial (z. B. als regelmäßiger Check-In in Schulklassen oder in digitalen Räumen, vgl. Love-Storm). Zudem sollten gemeinsam getragene Regeln erarbeitet werden, die den Umgang auch im digitalen Raum betreffen. Wichtig ist dabei eine Konsequenz in der Umsetzung und ggf. Sanktion bei Regelverstoß, um die Wirksamkeit der Regeln zu verdeutlichen, sowie eine Unterstützung von Personen in vulnerablen Situationen, um ein allgemeines Verständnis und Einhalten der Regeln zu sichern.

Die **Strukturen** beziehen sich schließlich auf die (sozialen) Medien selbst, die das Auftreten von Hass im Netz erst ermöglichen sowie die Rahmenbedingungen des Phänomens, die etwa durch die Gesetzgebung beeinflusst werden. Durch das Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (vgl. NetzDG) wurde in Deutschland ein sehr vielversprechender Rahmen geschaffen, um Hass im Netz schneller identifizieren und verfolgen zu können, was das Sicherheits- und Selbstwirksamkeitserleben von Betroffenen stärken, potenziell Ausübende abschrecken und Strafverfolgungsbehörden in ihrer Arbeit unterstützen soll. Zudem wurde

durch die Gründung von Initiativen gegen Hass im Netz (z. B. das stetig wachsense Kompetenznetzwerk gegen Hass im Netz, https://kompetenznetzwerk-hass-im-netz.de/) und die starke finanzielle Förderung der Netzwerkarbeit in diesem Feld eine einzigartige Struktur geschaffen, die das Potenzial hat, durch Medien- und Aufklärungsarbeit eine breite Öffentlichkeit zu erreichen, Multiplikator*innen auszubilden und relevante Lebenswelten anzusprechen. Dazu zählen Lern-, Entwicklungs- und Sozialisationskontexte wie Familie, Freizeit, Schule, Arbeit oder digitale Gruppen. Relevante Personen sind daher auch in allen Bereichen zu finden, wie Eltern, Lehrkräfte, Peers, digitale Freiwillige (z. B. Community Managers), Sozialpädagog*innen und Beschäftigte der Jugendarbeit (z. B. in Vereinen). Wenngleich sich ein großer Teil bisheriger Bemühungen an Betroffene, Kinder und Jugendliche und z. T. pädagogisches Fachpersonal und auch Eltern richtet, können weitere Akteure, etwa der Strafverfolgung oder der sozialen Arbeit in Zukunft stärker Berücksichtigung finden, um ganzheitliche Strategien zu entwickeln (vgl. Communities That Care-Ansatz; Kuklinski et al., 2021; Walter et al., 2023).

Da Personen, die in Behörden der Strafverfolgung tätig sind, selbst von Hass und Hetze betroffen sein können und vor der Herausforderung stehen, freie Meinungsäußerung, Hassphänomene und Straftatbestände voneinander in der Praxis abzugrenzen, sind auch sie eine wichtige Zielgruppe für strukturelle Fort- und Weiterbildung, die etwa Wissen zum Thema vermittelt, Schnittstellen zu anderen Bereichen herausstellt, Handlungskompetenzen stärkt und ein gemeinsames Vorgehen zur Prävention fördert. Eine Bedarfsanalyse in

diesen Bereichen wäre dafür ein wichtiges Instrument, um bedarfsgerecht agieren zu können.

Die Anbieter*innen sozialer Medien sind ebenfalls zur Kontrolle und Auskunft verpflichtet und können durch nutzerorientierte Kommunikation dazu beitragen, Hass im Netz niedrigschwellig zu adressieren. Gleichwohl ist die Umsetzung im Alltag ausbaufähig, da die Bekanntheit von Anlauf- und Meldestellen sowie die Kommunikation zum Thema mit Anbieter*innen und die Integration multipler Interventionen (z. B. Verhaltens- und Verhältnisebene) noch ausbaufähig ist.

Zuletzt ist erforderlich, **intersektorale Zusammenarbeit** zu stärken, zwischen Forschung und Entwicklung (z. B. Algorithmenentwicklung und -erprobung, sozialwissenschaftliche Validierung), Prävention und Bildungsarbeit (z. B. Medienbildung, schulische Prävention, Jugendarbeit) sowie Polizei und Justizbehörden (z. B. zur Identifikation von Hass im Netz und Strafverfolgung) und weiteren kommunalen Akteuren und Medien (z. B. Öffentlichkeitsarbeit). Auf diese Weise kann im Prozess gezielt interveniert werden. Hier sind neben der **Entwicklung von Netzwerken und Impulsen der Förderung praxisnaher Forschung auch Angebote der Fort- und Weiterbildung** sinnvoll, begleitet durch Evaluation zur Qualitätssicherung, um den aktuellen Erkenntnisstand zu reflektieren.

Dabei sind besonders **Schlüsselpositionen in Institutionen** entsprechend zu qualifizieren, um bedarfsgerecht und evidenzbasiert Entscheidungen treffen zu können und Empfehlungen geben zu können. Vergleichbar mit Modellen der kommunalen Gesundheitsförderung und (Kriminalitäts-)Prävention (z. B.

Communities That Care; Walter et al., 2023), sind auch in diesem Themenfeld Ansätze denkbar, die sektorenübergreifend Netzwerke aus relevanten, kommunalen Akteuren bilden und entsprechend qualifizieren, sodass ein gemeinsames Verständnis des Themenfelds und der Evidenzbasierung entstehen kann, partizipativ Strategien entwickelt werden können und daraus abgeleitete Maßnahmen lokal implementiert und evaluiert werden können, um den Stand der Forschung mit den örtlichen Prioritäten und Bedingungen zusammenzubringen.

Wenn Hass im Netz auch als Public Health-Problem verstanden wird, können diese Bemühungen zusammengeführt werden, sodass diese Aktivitäten der Prävention z. T. in bestehende Strukturen integriert werden kann – einige Programme mit guter Evidenzbasierung sind auch bereits Teil der Grünen Liste Prävention (z. B. Fairplayer, Medienhelden, Surf-Fair), die zu diesem Zweck im Communities That Care-Modell Anwendung findet. Wichtig ist dabei eine nachhaltige Finanzierung dieser Strukturen sowie der fortwährenden Qualifikation und Qualitätssicherung, um auch angesichts technologischer Innovation und der Dynamik sozialer Medien fortwährend handlungsfähig bleiben zu können.

Offene Forschungsfragen zum Themenfeld betreffen Aspekte der Entwicklung, Erreichung, Implementation und Evaluation: Fragen zur Entwicklung und Einbindung von digitalen Methoden und Interventionen betreffen die Potenziale und Risiken kombinierter und interaktiver Technologien in der Intervention (z. B. Mobile Sensing, Integration künstlicher Intelligenz, Just-In-Time-Adaptive-Interventions) sowie die

inter- und transdisziplinäre Zusammenarbeit, etwa zur Gestaltung, Annotierung und Teilung von (Trainings-)Datensätzen für maschinelles Lernen (vgl. Baer, 2019; Castaño-Pulgarín et al., 2021). Daran angeschlossen ist die Erreichung und Repräsentativität zu klären: Da die Forschung bisher vorrangig den Kinder- und Jugendbereich betrachtet hat, was angesichts der in der Gruppe stark ausgeprägten Mediennutzung verständlich ist, wurden weitere Gruppen von Betroffenen (z. B. Personen des öffentlichen Lebens, Personen in Justizbehörden) bislang deutlich seltener untersucht, sodass zu klären ist, inwiefern Prozesse, Risiko- und Schutzfaktoren übertragbar sind.

Die Vielzahl evidenzbasierter Interventionen und theoriebasierter Angebote stellt Fragen an die Implementation, etwa zur Akzeptanz, Wirksamkeit und Erreichbarkeit im Alltag und in verschiedenen Settings über die Lebensspanne. Mit Blick auf die Evaluation ist die geringe Evidenzbasierung bestimmter Programme, z. B. im Bereich Hassrede, hervorzuheben und eine Forderung nach komplexer Evaluation in Lebenswelten zu betonen (sog. *Effectiveness*).

Insgesamt kann festgehalten werden, dass für die Phänomene digitaler Hass und digitale Hetze ein hohes Interesse besteht und in Deutschland vielfältige Ressourcen für die Aufklärung und Netzwerkarbeit bereitgestellt werden. Auch in Niedersachsen sind wesentliche Strukturen bereits implementiert und es werden evidenzbasierte Programme zur Prävention und Intervention umgesetzt. Gleichwohl werden in der kritischen Würdigung der Situation zentrale Schwachpunkte und Herausforderungen der Umsetzung deutlich und offene Fragen für die Präventionsforschung sichtbar, etwa mit Blick auf die

Implementation, Effektivität und Nachhaltigkeit unter Alltagsbedingungen, der Kontextualität des Geschehens und der Übertragbarkeit auf diverse Zielgruppen und (neue) Phänomene. Somit bestehen ein großer Handlungsspielraum und ein großes Potenzial für die zukünftige Bearbeitung von Hass im Netz.

Implikationen für die Forschung und Praxis

- Individuum: Empfohlen ist eine Stärkung sozio-emotionaler Kompetenzen, Medienkompetenz, Digital Citizenship, Selbstwert und Empathie
- Gruppen: Die Entstehung und Gestaltung positiver Gruppennormen (vgl. Love-Storm) sollte weiter untersucht werden, idealiter durch partizipative Forschung und Methoden
- Schnittstellen von Online- und Offline-Welten sollten genauer in den Blick genommen werden, unter Beachtung diverser Einflüsse (z. B. Influencer*innen, Eltern, Peers)
- Strukturen: Hass im Netz ist ganzheitlich zu betrachten (z. B. als Public Health-Problem), entsprechend sollten Strukturen der Kooperation etabliert werden (vgl. Communities That Care)
- Kompetenzen zur intersektoralen Zusammenarbeit, Prävention und Intervention sind durch Fort- und Weiterbildung sicherzustellen, um nachhaltig systemische Veränderung zu erzielen
- Weitere Studien sollten weitere Lebenswelten (z. B. Arbeit neben Familie und Schule), Zielgruppen (z. B. Behörden, Minoritäten) beachten, stärker partizipativ in die Forschung einbinden und längsschnittlich Wirksamkeit und Implementation unter Alltagsbedingungen untersuchen

Angaben zur Positionalität

Dieses Gutachten wurde leitend durch Samuel Tomczyk erstellt, der seit einigen Jahren zu Themen der digitalen Gesundheit, Sicherheit und Prävention forscht und sich mit Gelingensbedingungen und Herausforderungen der Entwicklung, Implementation und Evaluation digitaler und digital gestützter Anwendungen über die Lebensspanne befasst. Als männlich gelesene, weiße Person, die in einem westlichen, industrialisierten Land forscht, das einen wesentlichen, historischen Bezug zu Hassphänomenen besitzt, ist die Perspektive auf den Gegenstand entsprechend kulturell und durch Sozialisation geprägt, sodass sich die hier vorgestellten Interpretationen und Schwerpunktsetzungen von Personen mit anderen Erfahrungs-, Wissens- und Erlebnishintergründen unterscheiden können. In der Erstellung des Gutachtens wurde sowohl in der Konzeption, Administration (z. B. im Forschungsteam) als auch in der theoretischen und empirischen Arbeit Wert auf Diversität der Perspektiven, Erfahrungen und Hintergründe gelegt, um mögliche Verzerrungen offenlegen und diskutieren zu können. Zugleich ist das Gutachten als eine Einladung zum Dialog zu verstehen, um weitere Diskussion anzuregen und die Prävention und Intervention bei Hass im Netz bestmöglich stärken zu können.

Literaturverzeichnis

Abarna, S., Sheeba, J. I., Jayasrilakshmi, S., & Devaneyan, S. P. (2022). Identification of cyber harassment and intention of target users on social media platforms. *Engineering Applications of Artificial Intelligence, 115*, 105283.

Ajzen, I. (2012). Martin Fishbein's legacy: The reasoned action approach. *The ANNALS of the American Academy of Political and Social Science, 640*(1), 11-27.

al-Utbi, M. I. K. (2019). A Critical Discourse Analysis of Hate Speech. *Journal of the College of Languages*, 39, 19-40.

Alhaboby, Z. A., Barnes, J., Evans, H., & Short, E. (2019). Cyber-victimization of people with chronic conditions and disabilities: a systematic review of scope and impact. *Trauma, Violence, & Abuse, 20*(3), 398-415.

Anderson, C. A., & Bushman, B. J. (2018). Media violence and the general aggression model. *Journal of Social Issues, 74*(2), 386-413.

Baacke, D. (1996). Medienkompetenz–Begrifflichkeit und sozialer Wandel. In A. v. Rein (Hg.), *Medienkompetenz als Schlüsselbegriff* (S. 112-124). DIE.

Baer, T. (2019). Algorithmic Biases and Social Media. *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists,* 95-106.

Baier, D., Hong, J. S., Kliem, S., & Bergmann, M. C. (2019). Consequences of bullying on adolescents' mental health in Germany: Comparing face-to-face bullying and cyberbullying. *Journal of Child and Family Studies, 28*, 2347-2357.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*(1), 1-26.

Bandura, A. (2011). Moral Disengagement. In D. J. Christie (Ed.), *The Encyclopedia of Peace Psychology* (pp. 1-5).

Banyard, V. L. (2011). Who will help prevent sexual violence: Creating an ecological model of bystander intervention. *Psychology of Violence, 1*(3), 216-229.

Beelmann, A. (2020). A social-developmental model of radicalization: A systematic integration of existing theories and empirical research. *International Journal of Conflict and Violence (IJCV)*, *14*, 1-14.

Bennett, S., Banyard, V. L., & Garnhart, L. (2014). To act or not to act, that is the question? Barriers and facilitators of bystander intervention. *Journal of Interpersonal Violence*, *29*(3), 476-496.

Bergmann, M. C. (2022). Comparing school-related risk factors of stereotypical bullying perpetration and cyberbullying perpetration. *European Journal of Criminology*, *19*(1), 77-97.

Bergmann, M. C., & Baier, D. (2018). Prevalence and correlates of cyberbullying perpetration. Findings from a German representative student survey. *International Journal of Environmental Research and Public Health*, *15*(2), 274.

Bergmann, M. C., Dreißigacker, A., von Skarczinski, B., & Wollinger, G. R. (2018). Cyber-dependent crime victimization: the same risk for everyone? *Cyberpsychology, Behavior, and Social Networking*, *21*(2), 84-90.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*, 3-33.

Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior*, *45*, 163-172.

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75-86.

Braveman, P. (2006). Health disparities and health equity: concepts and measurement. *Annual Review of Public Health*, *27*, 167-194.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513.

Bruns, A. (2019). It's not the technology, stupid: How the 'Echo Chamber'and 'Filter Bubble'metaphors have failed us. *International Association for Media and Communication Research*, 1-12.

Buelga, S., Cava, M.-J., Ruiz, D. M., & Ortega-Barón, J. (2022). Cyberbullying and suicidal behavior in adolescent students: A systematic review. *Revista de Educación*, *397*, 43-66.

Bushman, B. J., & Anderson, C. A. (2002). Violent video games and hostile expectations: A test of the general aggression model. *Personality and Social Psychology Bulletin*, *28*(12), 1679-1686.

Campbell, M., Whiteford, C., & Hooijer, J. (2019). Teachers' and parents' understanding of traditional and cyberbullying. *Journal of School Violence*, *18*(3), 388-402.

Casale, S., Fiovaranti, G., & Caplan, S. (2015). Online Disinhibition. *Journal of Media Psychology*, *27*(4), 170-177.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608.

Chen, L., Ho, S. S., & Lwin, M. O. (2017). A meta-analysis of factors predicting cyberbullying perpetration and victimization: From the social cognitive and media effects approach. *New Media & Society*, *19*(8), 1194-1213.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108-118.

Ching, H., Daffern, M., & Thomas, S. (2012). Appetitive violence: A new phenomenon? *Psychiatry, Psychology and Law*, *19*(5), 745-763.

Ching, H., Daffern, M., & Thomas, S. (2017). A comparison of offending trajectories in violent youth according to violence type. *Criminal Behaviour and Mental Health*, *27*(1), 8-14.

Choi, M. (2016). A concept analysis of digital citizenship for democratic citizenship education in the internet age. *Theory & Research in Social Education*, *44*(4), 565-607.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: a routine activity approach. *American Sociologial Review*, *44*(4), 588-608.

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online

experiences of those who produce online hate materials. *Violence and Gender*, *5*(1), 55-60.

Cuevas, J. A., & Dawson, B. L. (2021). An integrated review of recent research on the relationships between religious belief, political ideology, authoritarianism, and prejudice. *Psychological Reports*, *124*(3), 977-1014.

Delgado, R., & Stefancic, J. (1998). Critical race theory: Past, present, and future. *Current Legal Problems*, *51*(1), 467.

Dölling, D. (2012). Menschenbilder in der Kriminologie. In M. Hilgert & M. Wink (Hrsg.), *Menschen-Bilder: Darstellungen des Humanen in der Wissenschaft* (S. 281-289).

Domínguez-Hernández, F., Bonell, L., & Martínez-González, A. (2018). A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *12*(4).

Doty, J. L., Girón, K., Mehari, K. R., Sharma, D., Smith, S. J., Su, Y.-W., . . . Rousso, B. (2021). The dosage, context, and modality of interventions to prevent cyberbullying perpetration and victimization: A systematic review. *Prevention Science*, 1-15.

Douglas, K. M., McGarty, C., Bliuc, A.-M., & Lala, G. (2005). Understanding cyberhate: Social competition and social creativity in online white supremacist groups. *Social Science Computer Review*, *23*(1), 68-76.

Dutkowska-Zuk, A., Hounsel, A., Morrill, A., Xiong, A., Chetty, M., & Feamster, N. (2022). How and why people use virtual private networks. *31st USENIX Security Symposium* (USENIX Security 22).

Easton, S. D. (2014). Masculine norms, disclosure, and childhood adversities predict long-term mental distress among men with histories of child sexual abuse. *Child Abuse and Neglect*, *38*(2), 243-251.

Evangelio, C., Rodriguez-Gonzalez, P., Fernandez-Rio, J., & Gonzalez-Villora, S. (2022). Cyberbullying in elementary and middle school students: A systematic review. *Computers & Education*, *176*, 104356.

Figà Talamanca, G., & Arfini, S. (2022). Through the newsfeed glass: Rethinking filter bubbles and Echo chambers. *Philosophy & Technology*, *35*(1), 20.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., . . . Kainbacher, M. (2011). The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517.

Fulantelli, G., Taibi, D., Scifo, L., Schwarze, V., & Eimler, S. C. (2022). Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: a systematic review. *Frontiers in Psychology*, *13*, 909299.

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), 3.

George, J. J., & Leidner, D. E. (2019). From clicktivism to hacktivism: Understanding digital activism. *Information and Organization*, *29*(3), 100249.

Graf, D., Yanagida, T., & Spiel, C. (2019). Through the magnifying glass: Empathy's differential role in preventing and promoting traditional and cyberbullying. *Computers in Human Behavior*, *96*, 186-195.

Guo, S. (2016). A meta-analysis of the predictors of cyberbullying perpetration and victimization. *Psychology in the Schools*, *53*(4), 432-453.

Harden, S. M., Smith, M. L., Ory, M. G., Smith-Ray, R. L., Estabrooks, P. A., & Glasgow, R. E. (2018). RE-AIM in clinical, community, and corporate settings: perspectives, strategies, and recommendations to enhance public health impact. *Frontiers in Public Health*, *6*, 71.

Henares-Montiel, J., Benítez-Hidalgo, V., Ruiz-Pérez, I., Pastor-Moreno, G., & Rodríguez-Barranco, M. (2022). Cyberbullying and associated factors in member countries of the European Union: a systematic review and meta-analysis of studies with representative population samples. *International Journal of Environmental Research and Public Health*, *19*(12), 7364.

Hintz, A., Dencik, L., & Wahl-Jorgensen, K. (2017). Digital citizenship and surveillance| digital citizenship and surveillance society—introduction. *International Journal of Communication*, *11*, 9.

Hintz, E. A., & Betts, T. (2022). Reddit in communication research: current status, future directions and best practices. *Annals of the International Communication Association*, *46*(2), 116-133.

Humphreys, A. (2018). Social media. In M. R. Solomon & T. M. Lowrey (Eds.), *The Routledge Companion to Consumer Behavior* (pp. 363-379). Routledge.

Hutson, E., Kelly, S., & Militello, L. K. (2018). Systematic review of cyberbullying interventions for youth and parents with implications for evidence-based practice. *Worldviews on Evidence-Based Nursing*, *15*(1), 72-79.

Imperato, C., Schneider, B. H., Caricati, L., Amichai-Hamburger, Y., & Mancini, T. (2021). Allport meets internet: A meta-analytical investigation of online intergroup contact and prejudice reduction. *International Journal of Intercultural Relations*, *81*, 131-141.

Interian, R., G. Marzo, R., Mendoza, I., & Ribeiro, C. C. (2023). Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research*, *30*(6), 3122-3158.

Jadambaa, A., Thomas, H. J., Scott, J. G., Graves, N., Brain, D., & Pacella, R. (2019). Prevalence of traditional bullying and cyberbullying among children and adolescents in Australia: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry*, *53*(9), 878-888.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 126232.

Jamison, A. M., Broniatowski, D. A., & Quinn, S. C. (2019). Malicious actors on Twitter: A guide for public health researchers. *American Journal of Public Health*, *109*(5), 688-692.

Jeong, S.-H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of Communication*, *62*(3), 454-472.

Jia, Y., & Schumann, S. (2023). Tackling Hate Speech Online: The Effect of Counter-speech on Subsequent Bystander Reactions. *Preprint (without peer review)*. https://doi.org/https://doi.org/10.33767/osf.io/9jmza

John, A., Glendenning, A. C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., . . . Hawton, K. (2018). Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *Journal of Medical Internet Research*, *20*(4), e9044.

Jung, C. W. (2023). Role of informal social control in predicting racist hate speech on online platforms: collective efficacy and the theory of planned behavior. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 507-518.

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*, *24*(4), 2598-2615.

Karran, E. L., Cashin, A. G., Barker, T., Boyd, M. A., Chiarotto, A., Dewidar, O., . . . Moseley, G. L. (2023). Using PROGRESS-plus to identify current approaches to the collection and reporting of equity-relevant data: a scoping review. *Journal of Clinical Epidemiology*, *163*, 70–78.

Kaur, P., Dhir, A., Tandon, A., Alzeiby, E. A., & Abohassan, A. A. (2021). A systematic literature review on cyberstalking. An analysis of past achievements and future promises. *Technological Forecasting and Social Change*, *163*, 120426.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713-3744.

Kokkinos, C. M., & Antoniadou, N. (2019). Cyber-bullying and cyber-victimization among undergraduate student teachers through the lens of the General Aggression Model. *Computers in Human Behavior*, *98*, 59-68.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073.

Krebs, P., Prochaska, J. O., & Rossi, J. S. (2010). A meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine*, *51*(3), 214-221.

Kress, G. (1990). Critical discourse analysis. *Annual Review of Applied Linguistics*, *11*, 84-99.

Krieg, Y., Rook, L., Beckmann, L., & Kliem, S. (2022). *Adolescents in Lower Saxony: results of the Lower Saxony survey 2019*. Kriminologisches Forschungsinstitut Niedersachsen.

Kuklinski, M. R., Oesterle, S., Briney, J. S., & Hawkins, J. D. (2021). Long-term impacts and benefit–cost analysis of the communities that care prevention system at age 23, 12 years after baseline. *Prevention Science*, *22*, 452-463.

Kwan, B. M., McGinnes, H. L., Ory, M. G., Estabrooks, P. A., Waxmonsky, J. A., & Glasgow, R. E. (2019). RE-AIM in the real world: use of the RE-AIM framework for program planning and evaluation in clinical and community settings. *Frontiers in Public Health*, *7*, 345.

Lan, M., Law, N., & Pan, Q. (2022). Effectiveness of anti-cyberbullying educational programs: A socio-ecologically grounded systematic review and meta-analysis. *Computers in Human Behavior*, *130*, 107200.

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, *10*(3), 215-221.

Leukfeldt, E. R., & Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior*, *37*(3), 263-280.

Li, C., Wang, P., Martin-Moratinos, M., Bella-Fernández, M., & Blasco-Fontecilla, H. (2022). Traditional bullying and cyberbullying in the digital age and its associated mental health problems in children and

adolescents: a meta-analysis. *European Child & Adolescent Psychiatry*, 1-15.

Li, Q., Luo, Y., Hao, Z., Smith, B., Guo, Y., & Tyrone, C. (2021). Risk factors of cyberbullying perpetration among school-aged children across 41 countries: A perspective of routine activity theory. *International Journal of Bullying Prevention*, *3*, 168-180.

Lo Cricchio, M. G., Garcia-Poole, C., te Brinke, L. W., Bianchi, D., & Menesini, E. (2021). Moral disengagement and cyberbullying involvement: A systematic review. *European Journal of Developmental Psychology*, *18*(2), 271-311.

Lozano-Blasco, R., Quilez-Robres, A., & Latorre-Cosculluela, C. (2023). Sex, age and cyber-victimization: A meta-analysis. *Computers in Human Behavior*, *139*, 107491.

Mair, J. S., & Mair, M. (2003). Violence prevention and control through environmental modifications. *Annual Review of Public Health*, *24*(1), 209-225.

Marín, A., Hoyos, O., & Sierra, A. (2019). Risks and protective factors related to cyberbullying among adolescents: A systematic review. *Psychologist Papers*, *40*(2), 109-124.

Martinez-Cao, C., Gomez, L. E., Alcedo, M. Á., & Monsalve, A. (2021). Systematic review of bullying and cyberbullying in young people with intellectual disability. *Education and Training in Autism and Developmental Disabilities*, *56*(1), 3-17.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205-224.

Menesini, E., Calussi, P., & Nocentini, A. (2012). Cyberbullying and traditional bullying: Unique, additive, and synergistic effects on psychological health symptoms. In Q. Li, D. Cross, & P. K. Smith (Eds.), *Cyberbullying in the global playground: Research from international perspectives* (pp. 245-262). Blackwell.

Miller, R. B., & Brickman, S. J. (2004). A model of future-oriented motivation and self-regulation. *Educational Psychology Review*, *16*, 9-33.

Mishna, F., Cook, C., Saini, M., Wu, M. J., & MacFadden, R. (2009). Interventions for children, youth, and parents to prevent and reduce cyber abuse. *Campbell Systematic Reviews*, *5*(1), i-54.

Mohseni, M. R. (2023). Motives of Online Hate Speech: Results from a Quota Sample Online Survey. *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 499-506.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS One*, *15*(8), e0237861.

Mula Falcón, J., & Cruz González, C. (2023). Effectiveness of cyberbullying prevention programmes on perpetration levels: a meta-analysis. *Revista Fuentes*, 12-25.

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, *9*, 88364-88376.

Nguyen, T. (2023). Merging public health and automated approaches to address online hate speech. *AI and Ethics*, 1-10.

Nickel, S., & von dem Knesebeck, O. (2020). Effectiveness of Community-Based Health Promotion Interventions in Urban Areas: A Systematic Review. *Journal of Community Health*, *45*(2), 419-434.

Nocera, T. R., Dahlen, E. R., Poor, A., Strowd, J., Dortch, A., & Van Overloop, E. C. (2022). Moral disengagement mechanisms predict cyber aggression among emerging adults. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *16*(1), 1.

Nouh, M., Nurse, J. R., & Goldsmith, M. (2019). Understanding the radical mind: Identifying signals to detect extremist content on twitter. 2019 *IEEE International Conference on Intelligence and Security Informatics* (ISI).

Obermaier, M. (2022). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, 14614448221125417.

Ossa, F. C., Jantzer, V., Neumayer, F., Eppelmann, L., Resch, F., & Kaess, M. (2023). Cyberbullying and school bullying are related to additive

adverse effects among adolescents. *Psychopathology*, *56*(1-2), 127-137.

Petras, I.-K., & Petermann, F. (2019). Übersicht zu Risikofaktoren für Cybermobbing-Viktimisierung im Kindes-und Jugendalter und Empfehlungen für die Präventionsarbeit. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *67*(4), 203-220.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751.

Polanin, J. R., Espelage, D. L., Grotpeter, J. K., Ingram, K., Michaelson, L., Spinney, E., . . . Robinson, L. (2022). A systematic review and meta-analysis of interventions to decrease cyberbullying perpetration and victimization. *Prevention Science*, *23*(3), 439-454.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, *55*, 477-523.

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society*, *7*(4), 20563051211052906.

Rudnicki, K., Vandebosch, H., Voué, P., & Poels, K. (2023). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, *42*(5), 527-544.

Runions, K. C., & Bak, M. (2015). Online moral disengagement, cyberbullying, and cyber-aggression. *Cyberpsychology, Behavior, and Social Networking*, *18*(7), 400-405.

Ryan, P., & Lauver, D. R. (2002). The Efficacy of Tailored Interventions. *Journal of Nursing Scholarship*, *34*(4), 331-337.

Sabella, R. A., Patchin, J. W., & Hinduja, S. (2013). Cyberbullying myths and realities. *Computers in Human Behavior*, *29*(6), 2703-2711.

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A

systematic review. *International Journal of Market Research*, *62*(5), 578-598.

Seemann-Herz, L., Kansok-Dusche, J., Dix, A., Wachs, S., Krause, N., Ballaschk, C., . . . Bilz, L. (2022). Schulbezogene Programme zum Umgang mit Hatespeech – Eine kriteriengeleitete Bestandsaufnahme. *Zeitschrift für Bildungsforschung*, *12*(3), 597-614.

Seidler, Z. E., Dawes, A. J., Rice, S. M., Oliffe, J. L., & Dhillon, H. M. (2016). The role of masculinity in men's help-seeking for depression: a systematic review. *Clinical Psychology Review*, *49*, 106-118.

Simon, H., Baha, B. Y., & Garba, E. J. (2022). Trends in machine learning on automatic detection of hate speech on social media platforms: A systematic review. *FUW Trends in Science & Technology Journal*, *7*(1), 001-016.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136-146.

Soral, W., Malinowska, K., & Bilewicz, M. (2022). The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. *Peace and Conflict: Journal of Peace Psychology*, *28*(3), 361-371.

Sponholz, L. (2020). Der Begriff" Hate Speech" in der deutschsprachigen Forschung: eine empirische Begriffsanalyse. *SWS-Rundschau*, *60*(1), 43-65.

Sponholz, L. (2021). Hass mit likes: hate speech als Kommunikationsform in den social media. In S. Wachs, B. Koch-Priewe, & A. Zick (Hrsg.), *Hate Speech-Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen* (S. 15-37). Springer.

Stevens, F., Nurse, J. R., & Arief, B. (2021). Cyber stalking, cyber harassment, and adult mental health: A systematic review. *Cyberpsychology, Behavior, and Social Networking*, *24*(6), 367-376.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7-24). Nelson-Hall.

Tomczyk, S., Marlinghaus, L., Schmidt, S., & Bartha, S. (2023). Best possible selves in times of crisis: randomized controlled trial of best possible self-interventions during the COVID-19 pandemic. *The Journal of Positive Psychology*, 1-11.

Tomczyk, S., Pielmann, D., & Schmidt, S. (2022). More Than a Glance: Investigating the Differential Efficacy of Radicalizing Graphical Cues with Right-Wing Messages. *European Journal on Criminal Policy and Research*, *28*(2), 245-267.

Valkenburg, P. M., Peter, J., & Walther, J. B. (2016). Media effects: Theory and research. *Annual Review of Psychology*, *67*, 315-338.

Vranjes, I., Erreygers, S., Vandebosch, H., Baillien, E., & De Witte, H. (2018). Patterns of cybervictimization and emotion regulation in adolescents and adults. *Aggressive Behavior*, *44*(6), 647-657.

Vrontis, D., Makrides, A., Christofi, M., & Thrassou, A. (2021). Social media influencer marketing: A systematic review, integrative framework and future research agenda. *International Journal of Consumer Studies*, *45*(4), 617-644.

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Krause, N., Ballaschk, C., & Kansok-Dusche, J. (2022). The online hate speech cycle of violence: Moderating effects of moral disengagement and empathy in the victim-to-perpetrator relationship. *Cyberpsychology, Behavior, and Social Networking*, *25*(4), 223-229.

Wachs, S., Costello, M., Wright, M. F., Flora, K., Daskalou, V., Maziridou, E., . . . Biswal, R. (2021). "DNT LET'EM H8 U!": Applying the routine activity framework to understand cyberhate victimization among adolescents across eight countries. *Computers & Education*, *160*, 104026.

Wachs, S., Krause, N., Wright, M. F., & Gámez-Guadix, M. (2023). Effects of the Prevention Program "HateLess. Together against Hatred" on Adolescents' Empathy, Self-efficacy, and Countering Hate Speech. *Journal of Youth and Adolescence*, *52*(6), 1115-1128.

Wachs, S., Wright, M. F., & Vazsonyi, A. T. (2019). Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behaviour and Mental Health*, *29*(3), 179-188.

Walter, U., Groeger-Roth, F., & Röding, D. (2023). Evidence-based prevention for child and adolescent mental health: the" Communities That Care"(CTC) approach for Germany. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *66*(7), 774-783.

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology*, *45*, 101298.

Waqas, A., Salminen, J., Jung, S.-g., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One*, *14*(9), e0222194.

Wegge, D., Vandebosch, H., Eggermont, S., & Pabian, S. (2016). Popularity through online harm: The longitudinal associations between cyberbullying and sociometric status in early adolescence. *The Journal of Early Adolescence*, *36*(1), 86-107.

Windisch, S., Wiedlitzka, S., Olaghere, A., & Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, *18*(2), e1243.

Wright, M. F., & Wachs, S. (2020). Adolescents' cyber victimization: The influence of technologies, gender, and gender stereotype traits. *International Journal of Environmental Research and Public Health*, *17*(4), 1293.

Zych, I., Baldry, A. C., Farrington, D. P., & Llorent, V. J. (2019). Are children involved in cyberbullying low on empathy? A systematic review and meta-analysis of research on empathy versus different cyberbullying roles. *Aggression and Violent Behavior*, *45*, 83-97.