Maßnahmen gegen Online-Hass(bilder)

Zur Governance von diskriminierenden, beleidigenden oder zu Gewalt aufrufenden (visuellen) Inhalten im Netz

Franziska Oehmer-Pedrazzi / Stefano Pedrazzi*

Der interdisziplinäre Beitrag widmet sich der Identifikation, Systematisierung und Diskussion von Maßnahmen gegen (visuellen) Hass im Netz. Er folgt dabei einem zweistufigen Prozess: Zunächst werden empirische kommunikationswissenschaftliche Erkenntnisse zu den Merkmalen des (visuellen) Hasses gewonnen und diskutiert. Diese Erkenntnisse bilden die Grundlage für die darauffolgende Identifikation relevanter Maßnahmen. Einer Governance-Perspektive folgend werden dabei Handlungsoptionen des Staates (unter besonderer Berücksichtigung des Strafrechts), von Organisationen (insbesondere Plattformen, aber auch publizistische Medien und Parteien) sowie Nutzenden berücksichtigt. Zudem werden präventive (Hassinhalte verhindernde) als auch repressive (Hassinhalte sanktionierende) Maßnahmen besprochen. Der Beitrag argumentiert, dass Maßnahmen nur im Zusammenspiel und unter gegenseitiger Kontrolle verschiedener Akteur:innen Hass online wirksam vermindern und gleichzeitig ein Höchstmaß an Meinungsäußerungsfreiheit garantieren können.

Schlüsselwörter: Hassrede, Hassbilder, Governance, Plattformregulierung

Countering Online Hate (Images)

On the Governance of Discriminatory, Offensive, or Violence-Inciting (Visual) Content on the Internet

This interdisciplinary contribution focuses on the identification, systematization, and critical discussion of measures to counter (visual) hate in online environments. The study follows a two-stage approach: first, it synthesizes empirical insights from communication science regarding the characteristics and dynamics of (visual) hate. These findings form the basis for an identification of effective countermeasures. Adopting a governance perspective, the article explores potential interventions by various actors, including the state (with particular emphasis on criminal law), organizations (especially digital platforms, but also media outlets and political parties), and individual users. Both preventive strategies—aimed at deterring the dissemination of hate content—and repressive strategies—focused on sanctioning such content—are systematically examined. The paper argues that coordinated interplay and mutual oversight among these diverse stakeholders are essential to effectively mitigate online hate while preserving the highest possible standards of freedom of expression.

Key words: Hate speech, hate images, governance, platform regulation

* Prof. Dr. Franziska Oehmer-Pedrazzi, Fachhochschule Graubünden, Institut für Multimedia Production (IMP), Holzikofenweg 8, 3007 Bern, Schweiz, und Mileva Institut für Digitales und Gesellschaft, Chutzenweg 8, 3125 Toffen, Schweiz, franziska.oehmer@fhgr.ch, https://orcid.org/0000-0003-4005-9659;
Dr. Stefano Pedrazzi, Université de Fribourg, Departement für Kommunikationswissenschaft und Medienforschung, Bd de Pérolles 90, 1700 Fribourg, Schweiz, stefano.pedrazzi@unifr.ch, https://orcid.org/0000-0002-1600-6023.

1. Einleitung und Zielstellung

In der Europäischen Union geben 80 Prozent der Befragten an, dass sie online auf Hassrede gestoßen sind. 40 Prozent haben sich dabei sogar selbst angegriffen und bedroht gefühlt (Gagliardone et al., 2015). In der Schweiz gaben 69 Prozent der Bevölkerung an, bereits Hassrede im Netz gesehen zu haben – ein Drittel kommt nach eigenen Angaben sogar häufig damit in Kontakt (Wirz & Blassnig, 2024). Kinder und Jugendliche sind, so zeigt eine Metastudie, je nach Studiendesign, Land und Zeitraum zwischen sieben und 23,4 Prozent Opfer von Hassrede (Kansok-Dusche et al., 2023).

Mit Schneiders (2021) kann Hassrede definiert werden als „öffentliche und intentionale Äußerungen […], die beleidigend, einschüchternd oder belästigend sind und/oder zu Gewalt, Hass oder Diskriminierung aufrufen". Die Äußerungen weisen dabei gruppenbezogene Aggressionen auf, d. h. Personen oder Personengruppen werden aufgrund bestimmter Merkmale wie bspw. Herkunft, Geschlecht, Religion oder sexueller Orientierung angegriffen. Ist ein solcher Gruppenbezug nicht erkennbar, dann kann es sich zwar um eine Beleidigung oder auch um eine Form des Cyberbullying handeln, aber nicht um Hassrede (Unger & Unger-Sirsch, 2023). Hass kann dabei in unterschiedlichen Intensitätsstufen (Baider, 2020; Paasch-Colberg et al., 2021) – von humorvoll über aggressiv bis hin zu gewaltverherrlichend und -propagierend – vermittelt werden. Letzteres gilt dabei in den meisten Ländern als illegitim und somit rechtlich sanktionierbar, während Ersteres meist durch die Meinungsäußerungsfreiheit geschützt wird.

Hass wird nicht nur als Text vermittelt, sondern – besonders wirkungsvoll – auch in visueller Form bspw. durch Memes oder Cartoons (Brison, 2025). Dies lässt sich erstens auf den Bedeutungszuwachs von global agierenden Plattformen und sozialen Medien wie Instagram oder YouTube, die die Verbreitung visueller Inhalte befördern, zurückführen (Marquart, 2023). Zweitens ist dies in den Charakteristika visueller Inhalte selbst begründet: Sie binden Aufmerksamkeit, sind in der Regel leicht verständlich und werden wahrscheinlicher erinnert (Carney & Levin, 2002; Knobloch et al., 2003). Drittens ermöglichen technische Entwicklungen (generative KI; Fotobearbeitungssoftware ...) ein schnelles und auch für Laien zugängliches Erstellen von visuellen Inhalten (George, 2014).

Hass kann dabei sowohl auf individueller als auch auf gesellschaftlicher Ebene zu negativen Konsequenzen führen: Für die Betroffenen ist Hassrede oft mit psychischem, sozialem, wirtschaftlichem und sogar körperlichem Leid verbunden (Stahel et al., 2022, S. 5; Unger & Unger-Sirsch, 2023). Besonders schädlich für pluralistische und demokratische Gesellschaften ist es, wenn sich die Opfer von Hassbotschaften aus Angst vor weiterer Feindseligkeit aus dem öffentlichen Raum zurückziehen (Stahel et al., 2022) oder sich nicht mehr als gleichwertige und -berechtigte Mitglieder einer Gesellschaft fühlen (Unger & Unger-Sirsch, 2023). Auf gesellschaftlicher Makroebene können Hassbotschaften ein Klima der Intoleranz und Angst schaffen (Stahel et al., 2022) und zu einem inzivilen Online-Diskurs beitragen (Del Vigna et al., 2017). Zudem lässt sich auch ein Zusammenhang zwischen dem Ausmaß von Hassrede und Hassverbrechen feststellen (Müller & Schwarz, 2018; 2020).

Aufgrund des nicht auf nationale Grenzen beschränkten Aktionsfelds von digitalem Hass und ihrer Absender:innen (Schünemann & Steiger, 2023) und der Vielzahl an involvierten Akteur:innen können Maßnahmen gegen Hass nur aus einer Governance-Perspektive und damit unter Berücksichtigung verschiedener Regulierungsebenen und dem Zusammenspiel verschiedener Akteure wie Gesetzgebern, sozialen Netzwerken und Plattformen sowie deren Nutzenden erarbeitet werden (Helberger et al., 2018).

Die aktuelle Forschung zur Governance von Hass im Netz lässt sich anhand von vier Merkmalen charakterisieren:

- Erstens konzentriert sie sich auf regulatorische Ansätze insbesondere im Kontext des Digital Services Act (DSA) oder des zuvor geltenden deutschen Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz) (Schünemann & Steiger, 2023; Rensinghoff, 2022).
- Zweitens liegt der Fokus auf der Analyse von Hassrede oder der Identifikation von Maßnahmen, die auf Social-Media-Plattformen implementiert werden können (Demus et al., 2023; Jaki, 2023).
- Drittens bleiben Besonderheiten von Bildinhalten, die eine zentrale Rolle spielen können, weitgehend unbeachtet bzw. unreflektiert. Dies kann auch darin begründet liegen, dass visueller Hass meist interpretations- und kontextabhängig und damit häufig schwerer zu identifizieren und entsprechend auch regulierbar ist (Brown, 2018).
- Zudem mangelt es, viertens, an einer hinreichenden Berücksichtigung der Einflussmöglichkeiten von Nutzenden von digitalen (Kommunikations-)Plattformen (Helberger et al., 2018).

Der vorliegende Beitrag möchte sich diesem Forschungsdesiderat widmen und Governance-Maßnahmen unter Berücksichtigung kommunikationswissenschaftlicher Evidenz, verschiedener Akteure sowie der Spezifika von visuellen Inhalten identifizieren und diskutieren. Er folgt dabei einem zweistufigen Prozess:

- Zunächst sollen empirische Erkenntnisse zu den Merkmalen des (visuellen) Hasses gewonnen werden. Im Zentrum stehen dabei die Fragen, welche Absender, Hassobjekte und Kanäle sich in der (visuellen) Hasskommunikation identifizieren lassen.
- Diese Erkenntnisse bilden die Grundlage resp. das „Handlungsvoraussetzungswissen" (Dreyer, 2018, S. 63) für die darauffolgende evidenzbasierte Identifikation relevanter Governance-Maßnahmen zur Bekämpfung von digitalem:n Hass(bildern).

2. Merkmale von (visuellem) Hass

Im Folgenden reflektieren und systematisieren wir den aktuellen Stand der Forschung zu Hassrede nach Informationen über den Absender und den Adressaten von Hassbotschaften sowie den Kanälen, über die Hassbotschaften verbreitet werden. Die Fokussierung auf diese drei Merkmalsbereiche ist der Annahme geschuldet, dass empirische Erkenntnisse darüber für die Identifikation von Governance-Maßnahmen von besonderer Relevanz sind. Dabei wird Literatur zu textlicher und visueller Hassrede berücksichtigt. Unbeachtet bleiben in dieser Aufarbeitung Studien, die sich vornehmlich dem automatisierten Erkennen von Hass(bildern) oder dem Phänomen theoretisch-konzeptionell widmen.

2.1 Absender

Die Forschung zeigt, dass Hassrede oft aus strategischen Motiven von meist einflussreichen (politischen, religiösen, populärkulturellen) Akteur:innen veröffentlicht wird, die von politischem Erfolg, Verteidigung des eigenen Wertekanons, Unterhaltungszwecken oder finanziellen Gewinnen angetrieben werden. Die Absender von Hass betrachten sich selbst aufgrund ihres sozialen Status, ihrer Machtposition, ihres Bildungsniveaus, ihrer ethnischen Zugehörigkeit oder ihrer nationalen Identität als überlegen (Rabab'ah et al., 2024). Die Hassrede kann durch koordinierte Netzwerke oder individuell verbreitet werden (Erjavec & Kovačič, 2012; Frischlich et al., 2023). In letzterem Fall zeigen Untersuchungen, dass nur ein kleiner Teil der Bevölkerung aktiv zur Verbreitung von Hassrede beiträgt: In Europa gaben drei Prozent der Jugendlichen und jungen Erwachsenen an, Hassrede zu veröffentlichen (Kaakinen et al., 2018). In der Schweiz räumten 6,2 Prozent ein, Hassrede durch Posts, Likes oder Shares innerhalb eines Jahres verbreitet zu haben (Stahel et al., 2022). Der tatsächliche

Anteil könnte jedoch höher sein, da viele Menschen die veröffentlichten Inhalte möglicherweise nicht als Hassrede wahrnehmen, insbesondere bei visuellen Formaten wie Memes. Auf individueller Ebene sind Merkmale wie männliches Geschlecht, vorurteilsbehaftete Weltanschauungen, politische Einstellungen (populistisch; rechtsextrem) und mangelnde Empathiefähigkeit mit einer verstärkten Verbreitung von Hassrede verbunden (Frischlich et al., 2023). Auch im politischen Diskurs – insbesondere vor Wahlkämpfen – wird verstärkt u. a. von führenden Politiker:innen auf Hassrede zurückgegriffen (für Wahlkampfreden von Donald Trump: Rabab'ah et al., 2024; für die Wahlen zum Deutschen Bundestag: Ruttloff et al., 2022). Dies ist vor allem mit Blick auf die Vorbildfunktion politischer Akteur:innen von Relevanz: So konnten Studien aufzeigen, dass die Verbreitung von Hassrede durch führende Politiker:innen zu einer Verrohung der sozialen Normen und damit Normalisierung von Hass beiträgt, der sich u. a. auch in einem Anstieg an Hassrede und Hassverbrechen manifestiert (Müller & Schwarz, 2018; 2020; Kim & Ogawa, 2024).

Eine Studie, die sich der Analyse von Hassbildern widmet (Oehmer-Pedrazzi & Pedrazzi, 2024), zeigt zudem, dass in einem Drittel der Fälle die Hassbilder aktiv von Organisationen und kollektiven Akteuren wie politischen Parteien oder Medienorganisationen und damit von Akteuren, die über finanzielle und personelle Ressourcen verfügen, verbreitet werden. Auch Einzelakteur:innen – darunter Politiker:innen – teilen Hassbilder von ihren persönlichen Accounts.

2.2 Objekt

Die Forschung zeigt, dass vor allem Mitglieder marginalisierter Gruppen von Online-Hassrede betroffen sind (Unger & Unger-Sirsch, 2023): Menschen werden aufgrund ihrer Religion (Horsti, 2017; Hanzelka & Schmidt, 2017; Farkas et al., 2018), ihrer Hautfarbe (Ben-David & Fernández, 2016), ihres Geschlechts oder ihrer sexuellen Orientierung (Sobieraj, 2018; Lillian, 2007) oder ihres Flüchtlingsstatus (Kreis, 2017; Merrill & Åkerlund, 2018; Rabab'ah et al., 2024) angegriffen. Zudem geben auch politische Ansichten oder Aktivitäten Anlass, Zielscheibe von Hass und Aggression zu werden (Wirz & Blassnig, 2024).

Am häufigsten von Hassbildern betroffen sind Ausländer:innen und Migrant:innen (Oehmer-Pedrazzi & Pedrazzi, 2024): Jedes vierte Hassbild richtete sich gegen Personen anderer Nationalitäten. 21 Prozent der Bilder zeigen Hass gegen Menschen aufgrund ihres Geschlechts oder ihrer Geschlechtsidentität. Besonders häufig sind dabei transgeschlechtliche Personen sowie Frauen Ziel von Angriffen (ebd.). Verschiedene Positionen zu aktuellen Themen wie Klimaschutz, dem Ukraine-Konflikt oder Impfungen waren ebenfalls Motive für die Verbreitung von Hassbildern gegen die gegnerische Seite (ebd.).

2.3 Kanäle

Studien zu Hassrede haben Hassbotschaften auf einer Vielzahl von Kanälen gefunden und analysiert. Die meisten Studien konzentrieren sich auf große Kommunikationsplattformen wie Twitter/X (Burnap & Williams, 2015), Facebook (Farkas et al., 2018; Merrill & Åkerlund, 2018), Instagram (Frischlich et al., 2020) und YouTube (Murthy & Sharma, 2019), die für die Datenerhebung vergleichsweise zugänglicher sind (oder waren). Zudem gelten sie aufgrund fehlender redaktioneller Gatekeeper und berufsethischer Standards als besonders prädestiniert für die Verbreitung von Hassbotschaften. Die meistverwendeten Kanäle zur Verbreitung von Hassbildern sind Twitter/X und Instagram (Oehmer-Pedrazzi & Pedrazzi, 2024). Seit der Übernahme von Twitter durch Elon Musk und den damit einhergehenden Lockerungen der Inhaltsmoderation ist ein Anstieg von Hassbotschaften auf Twitter/X zu beobachten (Hickey et al., 2023). Twitter/X- und Instagram-Plattformen haben

unterschiedliche primäre Zielgruppen: Instagram wird vor allem von jüngeren Menschen genutzt (Külling et al., 2022), während Twitter/X überwiegend von Menschen mittleren Alters verwendet wird (Udris et al., 2024). Hassbilder sind also nicht auf einen einzigen Diskursraum oder ein spezifisches Zielpublikum beschränkt.

Zudem ist auch die Verbreitung von Hass durch journalistische Medien (Harlow, 2015; Sponholz, 2018) oder in Kommentarsektionen journalistischer Medien dokumentiert (Boberg et al., 2018; Frischlich et al., 2019; Paasch-Colberg et al., 2021). Auch Hassbilder finden in journalistischen Medien eine Plattform (Oehmer-Pedrazzi & Pedrazzi, 2024): Zehn Prozent der Hassbilder wurden auf den Websites von Zeitungen, Fernsehsendern oder Online-Medien veröffentlicht. Obwohl die Botschaften dieser Hassbilder oft in begleitenden Artikeln kritisch hinterfragt werden, bietet ihre Veröffentlichung eine zusätzliche Reichweite. Auch Kommentare in digitalen Spielen (vgl. Breuer, 2017; Gabriel, 2020; Yang et al., 2024) oder auf Streaming-Diensten wie bspw. Twitch (vgl. Müller, 25.01.2024) enthalten Hass, der sich oft gegen Frauen richtet.

Weitere Plattformen wie Reddit, Amazon oder tutti.ch (ein Schweizer Kleinanzeigenportal) dienen ebenfalls als Verbreitungskanal für Hassbilder (Oehmer-Pedrazzi & Pedrazzi, 2024). Marktplatz-Plattformen, die in der Regel eine hohe Nutzerzahl aufweisen, stehen nicht im primären Fokus der Kommunikationsforschung oder der Regulierungsmaßnahmen politischer und zivilgesellschaftlicher Akteure. Die Verbreitung von Hass über (private) Kommunikationskanäle, beispielsweise Messaging-Dienste wie WhatsApp oder Telegram (Vergani et al., 2022), ist im Vergleich weniger zugänglich, könnte jedoch eine besonders wichtige Rolle bei der Verbreitung visueller Hassrede spielen, da diese als privater Raum wahrgenommen werden, der als geschützter gilt.

3. Zur Governance von (visuellem) Hass

Der Beitrag folgt einer Governance-Perspektive und trägt damit dem Umstand Rechnung, dass die Eindämmung von Hassrede nicht allein staatlicher Regulierung, sondern einer kollaborativen Beteiligung verschiedener involvierter Akteur:innen bedarf. Unter Governance wird mit Mayntz (2004, S. 66) „das Gesamt aller nebeneinander bestehenden Formen der kollektiven Regelung gesellschaftlicher Sachverhalte" verstanden. Damit erlaubt der Governance-Ansatz einerseits eine Berücksichtigung verschiedener Formen der Regulierung, die von reinen Marktmechanismen bis hin zu gesetzlich bindender Regulierung durch staatliche Behörden reichen und intermediäre Formen von Governance einschließen, wie die Selbstorganisation durch einzelne Unternehmen, kollektive Selbstregulierung durch Branchen oder Ko-Regulierung (Latzer et al., 2003). Andererseits erweitert die Governance-Perspektive den Fokus auf nichtstaatliche und nichtnationale Akteure und integriert eine Vielzahl politischer, wirtschaftlicher und zivilgesellschaftlicher Akteure, während sie die zunehmende Bedeutung internationaler und transnationaler Phänomene, die im Bereich der digital vermittelten Kommunikation vorzufinden ist, berücksichtigt. Insbesondere bietet sie auch die Möglichkeit, soziotechnische Arrangements in die Analyse von Regulierungsprozessen und -rahmen zu integrieren (Gillespie, 2018; Katzenbach, 2020), was angesichts der Bedeutung digitaler Infrastrukturen und Technologien wie Plattformen und Algorithmen im Kontext der Analyse von Governance-Maßnahmen zu digital vermittelter (visueller) Hassrede besonders angemessen erscheint. Ein Verständnis von Governance als „interactions taken to solve societal problems and create societal opportunities" (Kooiman & Bavinck, 2005, S. 17) ermöglicht ferner die subsumierte Betrachtung sowohl bereits implementierter als auch diskutierter oder potenziell denkbarer sowie nach dem Zeitpunkt des Eingriffs in präventive und repressive Maßnahmen.

Die Identifikation der Governance-Maßnahmen erfolgte anhand einer Metaanalyse nach dem Verfahren des propositionalen Inventars (Bonfadelli & Meier, 1984). Es wurde dabei Literatur der Kommunikations- und Rechtswissenschaft zum Thema Hassrede berücksichtigt. Die relevanten Publikationen wurden in einem ersten explorativen Verfahren anhand von Recherchen auf Google Scholar ermittelt und anschließend in einem Schneeballverfahren durch Verweise in den vorgefundenen Publikationen ergänzt. Der Suchbefehl enthielt die Begriffe „Hassrede" und „Hate Speech". Berücksichtigt wurden Beiträge, die im Titel und im Abstract bzw. in der Einleitung einen Bezug zu Maßnahmen gegen Hass enthielten und ab 2020 veröffentlicht wurden. Die so recherchierten Publikationen wurden nach Hinweisen zu Governance-Maßnahmen sowie deren jeweiligen Vor- und Nachteilen durchsucht. Die Governance-Maßnahmen werden systematisiert nach Instrumenten, die sich gegen den *Adressaten* richten, die das *Hassobjekt* schützen und die darauf abzielen, den *Kanal* bzw. die *Plattform* zu einem förderlichen Diskurs- und Nutzungsraum zu gestalten. Wichtig ist dabei anzumerken, dass diese Systematisierung nicht immer trennscharf ist. Dabei wird der Fokus auch auf die Anwendbarkeit der Governance-Maßnahmen für Spezifika visueller Inhalte gelegt, die, anders als Texte, meist einen erhöhten Interpretationsbedarf notwendig machen.

3.1 Governance-Maßnahmen gegen den Absender

Governance-Maßnahmen, die sich an den Absender bzw. die Quelle von (visuellen) Hassbotschaften wenden, verfolgen zwei Zielstellungen: Zum einen soll die Veröffentlichung von Hassbotschaften verhindert oder sanktioniert werden. Zum anderen sollen die Urheber:innen von Hassbotschaften oder die Betreiber:innen von Kommunikationskanälen, in denen Hassbotschaften zirkulieren, dazu bewegt werden, bereits verbreitete Hassbotschaften zu löschen, um eine weitere Verbreitung durch Nutzende entsprechender Kommunikationskanäle zu verhindern. Die Forschung zeigt (vgl. Kap. 2.1), dass Hass von ressourcenstarken Einzel- und Kollektivakteuren verbreitet wird. Dazu zählen Parteien, einzelne Politiker:innen sowie Einzelpersonen mit einer Präferenz für populistische, extreme Parteien und Ideologien.

Staatliche Maßnahmen gegen die Absender:innen von Hassbotschaften können nur bei normiert illegalen Hassbotschaften greifen und sind trotz des internationalen Charakters von Online-Inhalten nach wie vor an territoriale Grenzen gebunden (Schünemann & Steiger, 2023). In vielen europäischen Staaten lassen sich jedoch ähnliche Regelungen für illegale Äußerungen und deren Sanktionierung finden. Meist werden dabei diskriminierende Aussagen geahndet, die öffentlich verbreitet werden und sich gegen – in der Regel abschließend benannte – Gruppierungen richten. In Deutschland wird beispielsweise mit § 130, Abs. 1 des Strafgesetzbuches „Volksverhetzung" bestraft, die sich gegen eine „nationale, rassische, religiöse oder durch ihre ethnische Herkunft bestimmte Gruppe" richtet. Die rechtlichen Rahmenbedingungen in der Schweiz, insbesondere Artikel 261bis der Strafnorm gegen „Diskriminierung und Hass" des schweizerischen Strafgesetzbuchs, erweitern seit 2020 den Kreis der Betroffenengruppe um Personen, die aufgrund ihrer sexuellen Orientierung herabgesetzt oder diskriminiert werden. Zudem ist es verboten, zu Gewalt aufzurufen (Artikel 259 Schweizer Strafgesetzbuch) oder Verbrechen gegen die Menschlichkeit zu rechtfertigen. Diskriminierungen aufgrund persönlicher Ansichten oder Status werden jedoch nicht abgedeckt. Zudem können diskriminierende oder herabsetzende Äußerungen als Ehrverletzungsdelikt verfolgt werden. Dies kann dann zur Anwendung kommen, wenn sich eine Aussage gegen eine bestimmte Person richtet (Rensinghoff, 2022). Bei der Beurteilung von diskriminierenden, herabsetzenden und ehrverletzenden Aussagen ist jeweils Satire oder die Kunstfreiheit als Rechtfertigungsgrund zu berücksichtigen (Schünemann & Steiger, 2023).

Zudem gilt jeweils im Einzelfall zwischen verfassungsrechtlich garantierter Meinungsäußerungsfreiheit und dem Recht auf Sanktionierung diskriminierender und herabwürdigender Inhalte abzuwägen (u. a. Ladeur & Gostomzyk, 2017; Schünemann & Steiger, 2023). Aus diesem Grund werden staatliche präventive Maßnahmen gegen den Absender von Hassbotschaften in modernen Demokratien als Zensur verneint.

Auf Ebene der Europäischen Union hat man 2022 mit dem Digital Services Act (DSA) dem Umstand Rechnung getragen, dass vor allem Plattformen für ihre jeweilige Diskurskultur selbst verantwortlich sind. Der DSA zielt darauf ab, ein sichereres und transparenteres digitales Umfeld durch die Regulierung von digitalen Diensten zu schaffen. Besonderes Augenmerk gilt dabei auch dem Schutz vor Hass: Digitale Dienste werden damit verpflichtet, illegale Inhalte schnellstmöglich zu löschen, sobald sie davon Kenntnis erlangen, entsprechende Melde- und Beschwerdeprozesse zu implementieren, ihre Moderationsprozesse transparent zu machen und dabei dem Schutz der Meinungsäußerungsfreiheit ein besonderes Gewicht einzuräumen. Besonders großen Plattformen wird darüber hinaus auferlegt, eine Risikobewertung über die potenziellen gesellschaftlichen Schäden ihrer Dienste (bspw. durch die Verbreitung von Hassbotschaften) vorzulegen, sich einem unabhängigen Audit zu unterziehen und mit unabhängigen Hinweisgebenden (Trusted Flaggers) zusammenzuarbeiten.

Plattformen haben verschiedene Möglichkeiten, Maßnahmen gegen die Verbreitung sowohl von illegalen als auch rechtlich möglichen und dennoch potenziell verletzenden Hassbotschaften zu ergreifen. Eine wesentliche präventive und repressive Maßnahme besteht darin, Hassbotschaften, die gegen rechtliche oder die Nutzungsrichtlinien der Plattformen selbst verstoßen, durch manuelle, automatisierte oder einer Kombination aus manueller und automatisierter Content Moderation an einer Veröffentlichung zu hindern oder sie als bedenklich zu kennzeichnen (Boberg et al., 2018; Gandhi et al., 2024). Aufgrund der Menge der zu begutachtenden Inhalte wäre eine alleinige manuelle Content Moderation nicht realisierbar (Gandhi et al., 2024). Kritisch anzumerken ist jedoch, dass die Identifikation von Hass und insbesondere Hassbildern erheblichem Interpretationsbedarf unterliegt, bei dem verschiedene nationale, historische und kulturelle Kontexte Berücksichtigung finden und entsprechend recherchiert und decodiert werden müssen. Das ist jedoch gerade vor dem Hintergrund beschränkter (zeitlicher) Ressourcen für die Content Moderator:innen kaum zu leisten. Automatisierte Erkennungsverfahren können hier ebenfalls (noch) nicht eine sichere Identifikation von Hass(bildern) garantieren. Das liegt auch darin begründet, dass bei Hassbildern bspw. in Form von Memes die negative Botschaft erst durch das Zusammenspiel von Text und Bild entstehen, die einzeln keinerlei bedenklichen Inhalt transportieren würden (Gandhi et al., 2024; Hermida & Santos, 2023; Oehmer-Pedrazzi & Pedrazzi, 2024). Auch die Weiterentwicklung der Sprache oder Kreation neuer Worte oder Wortverbindungen erschwert die automatisierte Identifikation von Hassbotschaften. Hassbotschaften, die auf subtile Art oder ironisch vermittelt werden, können ebenfalls nur schwer erkannt werden (Gandhi et al., 2024).

Ebenso zielführend kann es sein, präventiv die Monetarisierung von Inhalten, die Hassbotschaften enthalten, zu erschweren (Frischlich et al., 2023). Dies kann durch die Deaktivierung von Werbeeinnahmen für entsprechende Inhalte oder durch die Einschränkung des Zugangs zu monetären Funktionen beispielsweise für stark negative und emotionalisierende Inhalte erreicht werden. Eine weitere präventive Maßnahme besteht in einer an eine:n Urheber:in einer Hassbotschaft adressierte Warnung, dass bei Publikation eine Sperrung des Accounts erfolgt (Yildirim et al., 2023). Zusätzlich zu präventiven Ansätzen können Plattformen auch repressive Maßnahmen ergreifen, um gezielt gegen den Absender von Hassbotschaften vorzugehen. Neben der zeitlichen Suspendierung stellt die Löschung von

Accounts die einschneidendste Möglichkeit dar. Weitere repressive Maßnahmen umfassen die algorithmische Nicht-Priorisierung oder die Kennzeichnung von potenziell als Hassbotschaften identifizierten Inhalten, wodurch deren Verbreitung gezielt eingeschränkt werden soll. Darüber hinaus können Plattformen einfache Meldeoptionen einrichten, um Nutzenden die Möglichkeit zu geben, Verstöße schnell und effizient zu melden.

Neben Plattformen können auch *Parteien* als häufig identifizierte organisationale Quelle von Hassbotschaften einen Beitrag leisten, um Hass einzudämmen: Angesichts des Status und der einzigartigen Rolle dieser Akteure als Vertreter der Demokratie könnte die Diskussion darüber geführt werden, ob (möglicherweise gesetzliche oder selbst auferlegte[1]) Standards für die Kommunikation von politischen Parteien geschaffen werden sollten.

Zuletzt können *staatliche und zivilgesellschaftliche Akteure* gemeinsam einen Beitrag zur Eindämmung von Hassbotschaften leisten, indem Sie Bildungsangebote schaffen (bspw. klicksafe.de), die Nutzende bei der Erkennung von und im Umgang mit Hassbotschaften auf digitalen Kommunikationsplattformen unterstützen. Dabei gilt es insbesondere, auch für die Bedeutung visueller Kommunikation wie bspw. von Memes bei der Verbreitung von Hass zu sensibilisieren.

Tabelle 1: Übersicht Governance-Maßnahmen gegen den Absender

Governance	Staat	Organisationen	Nutzende
präventiv		*Plattformen:* Nutzungsbedingungen Content Moderation Installierung Meldeverfahren Monetarisierung von Hassbotschaften erschweren Suspendierungs-Warnung *Parteien:* Selbstverpflichtung zu hassfreier politischer Kommunikation	
	Staatliche und zivilgesellschaftliche Akteure: Bildungsangebote zur Erkennung von und zum Umgang mit Hassbotschaften		
repressiv	*Sanktionen gegen öffentliche/n* Diskriminierung oder Herabsetzung definierter Gruppen Aufruf zur Gewalt ehrverletzende Äußerungen	*Plattformen:* Suspendierung oder Löschung Account Algorithmische Depriorisierung potenzieller Hassbotschaften Kennzeichnung potenzieller Hassbotschaften	Hassbotschaft melden Account blockieren/ entfolgen bei widerrechtlicher Hassbotschaft Anzeige erstatten Gegenrede (Gegenbild)

1 Beispiele für selbstauferlegte Standards bilden das Fairness-Abkommen anlässlich der Bundestagswahlen 2025 oder der KI-Kodex anlässlich der Parlamentswahlen 2023 in der Schweiz.

Nutzende können durch ihr Verhalten auf der Plattform zur Sanktionierung möglicher Hassquellen beitragen und so durch freiwillige Mitwirkung dem Recht zur Geltung und Durchsetzung verhelfen (Unger & Unger-Sirsch, 2023): Zum einen können sie aktiv Gebrauch machen von den Optionen, die ihnen von den Plattformen zur Verfügung gestellt werden. Dazu zählen Meldeverfahren gegen beleidigende, diskriminierende oder zu Gewalt aufrufende Inhalte oder Accounts, die Hassbotschaften verbreiten, sowie das Blockieren oder Entfolgen solcher Accounts. Zum anderen haben sie die Möglichkeit, direkt Anzeige bei staatlichen Strafverfolgungsbehörden zu erstatten, um den Inhalt auf Widerrechtlichkeit prüfen zu lassen. Ferner haben sie die Option, dem Absender auch bei sachlichen oder humoristischen Formen der Hassrede durch Gegenrede, bspw. in Form eines Kommentars, zu begegnen (Hangartner et al., 2021). Für visuelle Inhalte könnte die Gegenrede auch in visueller Form als Gegenbild erfolgen. Es sollte jedoch berücksichtigt werden, dass Studien belegen, dass Mediennutzende Hasskommentare nicht immer (im selben Maß) wahrnehmen und entsprechend auch divergierend reagieren können (Schmid et al., 2022). Zudem wurden Hinweise gefunden, die aufzeigen, dass humoristische Memes zu einer Normalisierung von Hass verbreitenden Inhalten beitragen und sie demzufolge seltener als anzeigepflichtig wahrgenommen werden (vgl. Schmid, 2023).

3.2 Governance-Maßnahmen zum Schutz der Adressaten von Hass

Neben den Maßnahmen, die vor allem die Urherber:innen von Hassbotschaften in den Blick nehmen, müssen Governance-Maßnahmen auch den Schutz der Adressat:innen von Hassbotschaften in den Fokus rücken. Empirische Studien (vgl. Kap. 2.2) machen deutlich, dass sich Hass insbesondere gegen Ausländer:innen und Migrant:innen sowie gegen Personen aufgrund ihres Geschlechts und ihrer Geschlechtsidentität richtet. Auch Einstellungen zum aktuellen Zeitgeschehen (wie bspw. zum Ukrainekrieg, zum Klimawandel) geben der jeweils anderen Seite Anlass zur diskriminierenden, beleidigenden oder zu Gewalt aufrufenden Kommunikation.

Der *Staat* kann zum Schutz von Opfern von (digitaler) Hassrede verschiedene repressive Maßnahmen ergreifen. Eine bedeutende Maßnahme besteht darin, das Schutzobjekt staatlicher Regulierung zu erweitern, um möglichst viele Personen vor Hass zu schützen: In Deutschland könnte beispielsweise der Tatbestand der Volksverhetzung nicht nur bei öffentlichen Äußerungen gegen Personen aufgrund von Rasse, Geschlecht oder Religion angerufen werden, sondern auch aufgrund der sexuellen Identität. Darüber hinaus könnte der rechtliche Schutz auch auf weitere Formen der Diskriminierung ausgedehnt werden, die bisher möglicherweise nicht umfassend berücksichtigt wurden. Dies könnte etwa Einstellungen oder Weltanschauungen betreffen, die häufig Ziel von Hassrede sind, jedoch nicht unter die klassischen Diskriminierungsmerkmale fallen.

Da Hassbotschaften und insbesondere in visueller Form aufbereiteter Hass häufig in hohem Maße kontextabhängig sind, ist das Entschlüsseln der Botschaft oft auf kontextuelles Wissen angewiesen, das von kultureller, nationaler, sexueller Identität und der Zeit abhängig ist (Wilson & Land, 2020). Anders als Texten fehlt es insbesondere visuellen Inhalten in der Regel an identifizierbaren Begriffen oder Aussagen, die eine leichte Deutung ermöglichen könnten. Dies stellt eine besondere Herausforderung für die Inhaltsmoderation dar. Es scheint daher notwendig, dass die manuelle Moderation auf *Plattformen* von Personen durchgeführt wird, die unterschiedliche Altersgruppen, soziodemografische Merkmale und kulturelle Hintergründe repräsentieren. Dies wird besonders relevant, da Hass(bilder) als globales Problem betrachtet werden muss/müssen. Aufgrund des öffentlichen Charakters der über Plattformen vermittelten Kommunikation und der Verantwortung der Plattformen für diese Vorgänge ließe sich darüber hinaus ein Anspruch auf Kostenbeteiligung für Leis-

Tabelle 2: Übersicht Governance-Maßnahmen zum Schutz des Hassobjektes

Governance	Staat	Organisationen	Nutzende
präventiv		*Plattformen:* Diversität von Content Moderator:innen (v. a. bei der Prüfung visueller Inhalte)	Account auf der eigenen Timeline sperren; Nutzereinstellungen, z. B. Kommentarfunktion bei eigenen Posts ausschalten/ einschränken
	Staatliche und zivilgesellschaftliche Akteure: Bildungsangebote zur Erkennung von und zum Umgang mit Hassbotschaften		
repressiv	Schutzobjekte staatlicher Regulierung erweitern (bspw. in Deutschland: Schutz aufgrund sexueller Identität gewähren oder auch Diskriminierung aufgrund von Einstellungen mitberücksichtigen)	*Plattformen:* finanzielle Beteiligung an Opferhilfe	Solidarität durch Gegenrede

tungen zur Bewältigung negativer externer Effekte, wie z. B. Opferberatungsleistungen, ableiten (Das NETTZ et al., 2024).

Nutzende verfügen ebenfalls über Möglichkeiten, präventive Maßnahmen zum Schutz von Opfern von Hassrede im Netz zu ergreifen und damit auch die eigene Online-Interaktion sicherer zu gestalten. Eine zentrale Maßnahme besteht darin, Accounts, die Hassbotschaften verbreiten, auf der eigenen Timeline zu sperren. Durch das Sperren solcher Accounts wird verhindert, dass diese weiterhin in den eigenen sozialen Netzwerken sichtbar sind oder direkten Kontakt aufnehmen können. Dies trägt nicht nur zur Reduzierung der persönlichen Angriffsfläche bei, sondern kann auch eine präventive Wirkung entfalten, indem toxische Interaktionen frühzeitig unterbunden werden. Weiter ist die Einschränkung oder vollständige Deaktivierung der Kommentarfunktion bei eigenen Posts zu erwähnen. Indem Nutzende die Kommentarfunktion gezielt ausschalten, entziehen sie potenziellen Angreifenden die Möglichkeit, eigene Beiträge als Plattform für Hassrede zu nutzen. Alternativ können sie die Kommentarfunktion einschränken, etwa indem sie nur bestimmten Personen das Kommentieren erlauben. Bei den repressiven Maßnahmen können Bekundungen der Solidarität und Empathie mit den Adressaten von Hass mittels Gegenrede genannt werden (Hangartner et al., 2021). Die Praxis der Gegenrede kann zudem eine präventive Wirkung entfalten, da sie zeigt, dass Hassrede in der digitalen Öffentlichkeit nicht toleriert wird.

3.3 Governance-Maßnahmen für Plattformen und soziale Netzwerke

Aufgrund der zentralen Position, die Plattformen und sozialen Netzwerken in der Vermittlung von Hassbotschaften zukommt, würden Governance-Maßnahmen, die nur den:ie Urheber:in und die Adressaten berücksichtigen, zu kurz greifen. Daher müssen auch Governance-Maßnahmen diskutiert werden, die darauf abzielen, auf Plattformen und sozialen Netzwerken eine für die Diskurskultur und das Nutzungsverhalten positive Umgebung zu schaffen. Dies wird erreicht, indem u. a. die Publikation von Hassbotschaften reduziert oder sogar verhindert wird sowie deren Verbreitung und Reichweite eingeschränkt werden.

Tabelle 3: Übersicht Governance-Maßnahmen für Plattformen und soziale Netzwerke

Governance	Staat	Organisationen	Nutzende
präventiv		*Plattformen:* Content Moderation Governance by Design (u.a. Netzwerkarchitektur, algorithmische Selektion, Meldeprozess für Nutzende) *Publizistische Medien:* Content Moderation Selbstverpflichtung zum sensiblen Umgang mit Hass(botschaften)	
repressiv	Fokus nicht auf große Kommunikationsplattformen beschränken	*Publizistische Medien:* Moderation der Kommentarspalten	Wahl der Plattform unter Berücksichtigung der jeweiligen Diskurs- und Handlungskultur

Wie die Forschung deutlich macht (vgl. Kap. 2.3), werden Hassbotschaften allerdings nicht nur in sozialen Netzwerken verbreitet, sondern sind auch in Kanälen präsent, die bisher weniger stark im Fokus von Governance-Bemühungen standen: Messengerdienste, Gaming- und Livestreaming-Plattformen, Verkaufsportale sowie publizistische Medien (und ihre Kommentarspalten).

Aus Governance-Sicht deuten diese Erkenntnisse darauf hin, dass staatliche Maßnahmen, die sich ausschließlich auf Kommunikationsplattformen konzentrieren – wie derzeit bspw. in der Schweiz diskutiert (Bundesrat, 5. April 2023) – möglicherweise zu kurz greifen. Auch kommerzielle Plattformen wie Amazon, Kleinanzeigenportale sowie Computer- und Livestreaming-Plattformen bieten Hass einen Kanal. Ein Teil der auf EU-Ebene beschlossenen Maßnahmen im Rahmen des Digital Services Acts (DSA) inkludieren alle Plattformen, auf denen Hass zirkuliert. Jedoch werden nur besonders große Plattformen zu weiterführenden Maßnahmen verpflichtet (siehe Kap. 3.1). Plattformen mit geringeren Nutzendenzahlen, die dennoch von erheblicher Bedeutung in einem Land sind, wie Kleinanzeigenportale, bleiben bei bestimmten Instrumenten unberücksichtigt. Es wird angeregt, dass Maßnahmen zur Governance von (visuellem) Hass universell und damit reichweiten- und funktionsunabhängig greifen müssen. Plattformen, auf denen ein anonymes und disperses Publikum agieren kann, müssen in Governance-Maßnahmen vollumfänglich mitgedacht werden. Einzig Plattformen, die allein der privaten Kommunikation dienen, wären hiervon ausgenommen.

Plattformen selbst tragen maßgeblich zu ihrer eigenen Diskurs- und Handlungskultur bei: Sie bestimmen – bspw. über die Höhe der zur Verfügung gestellten personellen und finanziellen Ressourcen – über die Effektivität von Content Moderation und damit über das mögliche Filtern oder Markieren von (potenziell) diskriminierenden, beleidigenden oder zu Gewalt aufrufenden Inhalten. Zudem ermöglichen sie – durch ihre soziotechnischen Arrangements wie die Netzwerkarchitektur, algorithmische Selektionslogiken oder einer Meldefunktion für Nutzende –, ob und wie diese Inhalte geprüft, verbreitet oder gemeldet werden können.

Eine wesentliche Maßnahme, die sowohl von Plattformen als auch von publizistischen Medien umgesetzt werden kann, ist Content Moderation. Durch eine strenge Content

Moderation können Medien und Plattformen verhindern, dass diskriminierende oder beleidigende Inhalte verbreitet werden.

Publizistische Medien können zudem durch eine Selbstverpflichtung zum sensiblen Umgang mit Hassbotschaften einen zusätzlichen Schutz bieten. Diese Selbstverpflichtung könnte beinhalten, dass Medien in ihrer Berichterstattung verantwortungsvoll mit Hassrede umgehen, etwa durch sorgfältige Wortwahl, differenzierte Darstellungen und das Vermeiden von Sensationalismus, der die Sichtbarkeit von Hass verstärken könnte. Darüber hinaus können sie bewusst darauf verzichten, Hassrede und insbesondere Hassbildern eine Plattform zu bieten, indem sie diese nicht (unreflektiert) reproduzieren oder verbreiten.

Die *Nutzenden* könnten ihre Entscheidung für oder gegen eine Plattform auch von deren jeweiliger Diskurs- und Handlungskultur abhängig machen, wie dies beispielsweise bei X (ehemals Twitter) nach der Übernahme durch Elon Musk der Fall war, der insbesondere bei der Content Moderation Sparpotenzial geltend machte (Walser, 17.01.2024). Dies ist vor allem dann möglich, wenn Alternativen bestehen.

4. Fazit

Ziel des vorliegenden Beitrags war es, Governance-Maßnahmen von (visuellem) Hass zu identifizieren, zu systematisieren und zu diskutieren. Er folgte dabei einem zweistufigen Prozess: Zunächst wurden empirische Erkenntnisse zu den Merkmalen des (visuellen) Hasses gewonnen und diskutiert: Organisationen (vor allem Parteien) fungieren neben Einzelpersonen unter Klarnamen (darunter auch Politiker:innen) als Urheber:innen von Hassbotschaften. Betroffen von Hass(bildern) sind Gruppen aufgrund ihres Ausländer:innen- oder Migrant:innenstatus sowie ihrer Geschlechtsidentität. Hass zirkuliert auf verschiedenen Plattformen – vor allem auf jenen mit hoher Nutzendenbeteiligung ohne oder mit nur wenig inhaltsmoderierender Kontrolle.

Diese Erkenntnisse bildeten die Grundlage für die Identifikation relevanter präventiver und repressiver *Governance-Maßnahmen* zur Bekämpfung von digitalen Hass(bildern) unter Berücksichtigung kollaborativer und soziotechnischer Arrangements. Es konnte eine Vielzahl an bestehenden und potenziell möglichen Governance-Maßnahmen, die von staatlichen Akteur:innen, Organisationen sowie Nutzenden ergriffen werden können, dargelegt werden. Dazu zählen strafrechtliche Bestimmungen zum Schutz vor Diskriminierung, Content Moderation durch Plattformen sowie Maßnahmen zur Gegenrede durch Nutzende. Diskutiert wurden auch jeweils mögliche Herausforderungen und Risiken bei der Anwendung der Maßnahmen, wie bspw. die falsche (false positive) oder fehlende (false negative) Identifikation von Hassbotschaften durch manuelle und automatisierte Content Moderation der Plattformen (insbesondere bei multimodalen Inhalten), die verzerrte Beurteilung der Hassäußerungen durch Nutzende oder die Beschränkung staatlicher Maßnahmen auf illegitime Inhalte. Dies verdeutlicht, dass Maßnahmen gegen Hass nicht nur einzelnen Akteur:innen auferlegt werden sollten, sondern dass nur im Zusammenspiel und unter gegenseitiger Kontrolle Hass online wirksam vermindert und gleichzeitig ein Höchstmaß an Meinungsäußerungsfreiheit garantiert werden kann.

Unter Berücksichtigung der empirischen Befunde zu den Hassabsender:innen und -adressat:innen sowie den zur Verbreitung genutzten Kanälen, können folgende Handlungsempfehlungen für die Governance vom Hass formuliert werden:

– *Staat:* Der Rechtsschutz gegen Diskriminierung oder vor Volksverhetzung sollte nicht nur für bestimmte Gruppierungen geltend gemacht werden könnten, sondern bei jeder Herabsetzung oder Benachteiligung von Personen aufgrund geteilter Merkmale angerufen werden können.

- *Plattformen:* Um dem hohen – meist kulturell, national oder historisch geprägten – Interpretationsbedarf zur Decodierung von (vor allem visuellem) Hass gerecht zu werden, sind möglichst viele und diverse Content Moderator:innen grundlegend.
- *Parteien:* Als wichtige Repräsentanten und Vorbilder demokratischer Prozesse sollten Parteien im Rahmen von Selbstverpflichtungserklärungen auf eine aggressive, beleidigende oder zu Gewalt aufrufende Kommunikation insbesondere in der visuellen Kampagnenkommunikation innerhalb und außerhalb von Wahlkampf- und Abstimmungszeiten verzichten.
- *Medien:* Journalistische Medien sollten sich ihrer Rolle als mögliche reichweitenstarke Multiplikator:innen bei der Berichterstattung über Hassbotschaften bewusst werden und im Rahmen von Selbstverpflichtungen einen zurückhaltenden und sensiblen Umgang mit Hass anstreben, bspw. Hassbilder nicht zu veröffentlichen, sondern nur textlich zu beschreiben.

Die Vielzahl an möglichen Governance-Ansätzen muss jedoch, um Wirksamkeit entfalten zu können, nicht nur erkannt, sondern auch angewandt und umgesetzt werden: Dass empathische Gegenrede Personen davon abhalten kann, (weitere) Hassbotschaften zu verbreiten, setzt beispielsweise voraus, dass man sich als Gegenredner:in entsprechend auch öffentlich exponiert. Das kann mit Blick auf die nicht immer einzuschätzende Gegenreaktion des:r Urheber:s:in oder anderer Befürwortenden nicht immer erwartet werden. Content Moderation durch möglichst diverse Teams bedingt, dass die Plattformen beispielsweise auch entsprechende Ressourcen zur Verfügung stellen.

Alle Governance-Maßnahmen bedürfen zudem stets einer strengen Abwägung zwischen dem Recht auf freie Meinungsäußerung und dem Anspruch auf einen diskriminierungsfreien, inklusiven und gesunden öffentlichen Diskurs. Bei diesem Abwägungsprozess gilt es drei aktuellen Entwicklungen auf der Mikro- und Makroebene Rechnung zu tragen:

- Erstens können Hassbotschaften dazu führen, dass sich Personen aus dem öffentlichen Diskurs (Das NETTZ et al., 2024) oder sogar von öffentlichen Ämtern zurückziehen (Garne, 2022). Es ist entsprechend zu befürchten, dass in der öffentlichen Arena nicht mehr alle Stimmen Gehör finden. Ob auch in solchen Fällen das Recht auf freie Meinungsäußerung höher zu gewichten ist, muss geklärt werden.
- Zudem ist eine Governance von Hassrede auch, zweitens, verstärkt vor dem Hintergrund des wachsenden Erfolges von rechten oder sogar rechtsextremen Parteien gerade auch bei jungen Erwachsenen zu diskutieren, die es insbesondere verstehen, soziale Netzwerke kommunikativ – auch mit Hassbotschaften – zu nutzen (Vogel & Schmitt, 08.09.2024). Hier gilt es ebenfalls zu prüfen, welches Gewicht dem Recht auf freie Meinungsäußerung beizumessen ist, wenn es als möglicher Katalysator antidemokratischer Tendenzen wirkt.
- Drittens tragen Hassbotschaften auch dazu bei, der gesellschaftlichen Polarisierung Vorschub zu leisten (Romero-Rodríguez et al., 2023) – v. a. dann, wenn sich Hass gegen Personen mit anderen Einstellungen – bspw. zum Klimawandel oder zum Ukrainekrieg – richtet. Auch dies ist beim Abwägen zwischen dem Recht auf freie Meinungsäußerung sowie dem Anspruch auf einen diskursfreundlichen Raum zu berücksichtigen.

Dem übergeordnet ist jedoch auch zu klären, ob Maßnahmen gegen Hassbotschaften überhaupt einen Eingriff in die Meinungsäußerungsfreiheit darstellen, richten sie sich doch nur gegen die Form, und nicht gegen den Inhalt der Äußerung. Das Spektrum an Ansichten würde mit diesen Maßnahmen daher auch nicht beschnitten werden. Es würde lediglich eine bestimmte Art und Weise, wie eine Meinung dargelegt wird, unterbunden werden (vgl. Unger & Unger-Sirsch, 2023). Das beträfe sämtliche Hassintensitätsstufen und damit auch humorvolle Memes.

Literatur

Baider, F. H. (2020). Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society, 11*(2), 196–218. https://doi.org/10.1075/ps.20004.bai

Ben-David, A., & Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication, 10*, 1167–1193 .

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication, 6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Bonfadelli, H., & Meier, W. (1984). Meta-Forschung in der Publizistikwissenschaft. *Rundfunk und Fernsehen, 32*(4), 537–550.

Breuer, J. (2017). Hate speech in online games. In K. Kaspar, L. Gräßer, & A. Riffi (Eds.), *Online hate speech – Perspektiven auf eine neue Form des Hasses* (S. 107–112). kopaed.

Brison, S. J. (2025). Hate Speech. In *International Encyclopedia of Ethics*. Wiley. https://doi.org/10.1002/9781444367072.wbiee771.pub2

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities, 18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Bundesrat (05.04.2023). Grosse Kommunikationsplattformen: Bundesrat strebt Regulierung an. https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-94116.html [22.04.2025].

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 7*(2), 223–242.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*, 5–26.

Das NETTZ, Gesellschaft für Medienpädagogik und Kommunikationskultur, HateAid und Neue deutsche Medienmacher*innen als Teil des Kompetenznetzwerks gegen Hass im Netz (Hrsg.) (2024): Lauter Hass – leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Ergebnisse einer repräsentativen Befragung. Berlin. https://kompetenznetzwerk-hass-im-netz.de/download_lauterhass.php [22.04.2025].

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook, *Proceedings of the First Italian conference on cybersecurity,* 86–95).

Demus, C., Labudde, D., Pitz, J., Probol, N., Schütz, M., & Siegel, M. (2023). Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken. In S. Jaki & S. Steiger (Eds.), *Digitale Hate Speech* (S. 65–88). J. B. Metzler. https://doi.org/10.1007/978-3-662-65964-9_4

Dreyer, S. (2018). *Entscheidungen unter Ungewissheit im Jugendmedienschutz: Untersuchung der spielraumprägenden Faktoren gesetzgeberischer und behördlicher Entscheidungen mit Wissensdefiziten*. Nomos.

Erjavec, K., & Kovačič, M. P. (2012). „You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society, 15*(6), 899–920. https://doi.org/10/gfgnmm

Farkas, J., Schou, J., & Neumayer, C. (2018). Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society, 20*(5), 1850–1867. https://doi.org/10.1177/1461444817707759

Frischlich, L., Boberg, S., & Quandt, T. (2019). Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies, 20*(14), 2014–2033. https://doi.org/10/gfwcjc

Frischlich, L., Klapproth, J., & Brinkschulte, F. (2020). Between mainstream and alternative – Co-orientation in right-wing populist alternative news media. In C. Grimme, M. Preuß, F. W. Takes, & A. Waldherr (Eds.), *Disinformation in open online media* (pp. 150–167). Springer. https://doi.org/10.1007/978-3-030-39627-5_12

Frischlich, L., Schmid, U. K., & Rieger, D. (2023). Hass und Hetze im Netz. In M. Appel, F. Hutmacher, C. Mengelkamp, J. P. Stein, & S. Weber (Eds.), *Digital ist besser?! Psychologie der Online- und Mobilkommunikation* (S. 169–187). Springer. https://doi.org/10.1007/978-3-662-66608-1_14

Gabriel, S. (2020). Hate speech in der Computerspielkultur. In T. G. Rüdiger & P. Bayerl (Eds.), *Cyberkriminologie* (S. 217–230). Springer VS. https://doi.org/10.1007/978-3-658-28507-4_11

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., Poria, S., & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8), e13562. https://doi.org/10.1111/exsy.13562

Garne, J. (2022, 13. September). Prominente Abgänge im Kantonsrat: Sarah Akanji hört auf, weil sich die Angriffe auf ihre Person häufen. *Tages-Anzeiger*. https://www.tagesanzeiger.ch/sarah-akanji-hoert-auf-weil-sich-die-angriffe-auf-ihre-person-haeufen-194642477445 [22.04.2025].

George, C. (2014). Hate speech law and policy. In P. H. Ang and R. Mansell (Eds.), *The International Encyclopedia of Digital Communication and Society*. https://doi.org/10.1002/9781118767771.wbiedcs139

Gillespie, T. (2018). Governance of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE Handbook of Social Media* (pp. 254–278). SAGE Publications Ltd. https://doi.org/10.4135/9781473984066.n15

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., ... & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences, 118*(50).

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology, 11*(1), 143–160. https://doi.org/10.5281/zenodo.495778

Harlow, S. (2015). Story-chatters stirring up hate: Racist discourse in reader comments on U.S. newspaper websites. *Howard Journal of Communications, 26*(1), 21–42. https://doi.org/10.1080/10646175.2014.984795

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society, 34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Hermida, P. C. D. Q., & Santos, E. M. D. (2023). Detecting hate speech in memes: a review. *Artificial Intelligence Review*, *56*(11), 12833–12851. https://doi.org/10.1007/s10462-023-10459-7

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., & Burghardt, K. (2023). Auditing Elon Musk's impact on hate speech and bots. *Proceedings of the international AAAI conference on web and social media*. 17(1), 1133–1137. https://doi.org/10.1609/icwsm.v17i1.22222

Horsti, K. (2017). Digital islamophobia: The Swedish woman as a figure of pure and dangerous whiteness. *New Media & Society, 19*(9), 1440–1457. https://doi.org/10.1177/1461444816642169

Jaki, S. (2023). Hate speech in sozialen Medien: Ein Forschungsüberblick aus Sicht der Sprachwissenschaft. In S. Jaki & S. Steiger (Eds.), *Digitale Hate Speech: Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation* (S. 15–34). J. B. Metzler.

Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018). Social capital and online hate production: A four-country survey. *Crime, Law and Social Change, 69*(1), 25–39. https://doi.org/10.1007/s10611-017-9764-5

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. https://doi.org/10.1177/15248380221108070

Katzenbach, C. (2020). Die Governance sozialer Medien. In J.-H. Schmidt & M. Taddicken (Eds.), *Handbuch soziale Medien* (S. 1–24). Springer Fachmedien. https://doi.org/10.1007/978-3-658-03895-3_26-1

Kim, T., & Ogawa, Y. (2024). The impact of politicians' behaviors on hate speech spread: Hate speech adoption threshold on Twitter in Japan. *J Comput Soc Sc 7*, 1161–1186. https://doi.org/10.1007/s42001-024-00268-5

Kooiman, J., & Bavinck, M. (2005). The governance perspective. In J. Kooiman, M. Bavinck, S. Jentoft, & R. Pullin (Eds.), *Fish for life* (pp. 11–24). Amsterdam University Press. http://www.jstor.org/stable/j.ctt46mzgb.4

Knobloch, S., Hastall, M., Zillmann, D., & Callison, C. (2003). Imagery effects on the selective reading of Internet newsmagazines. *Communication Research, 30*(1), 3–29.

Kreis, R. (2017). #refugeesnotwelcome: Anti-refugee discourse on Twitter. *Discourse & Communication, 11*(5), 498–514. https://doi.org/10.1177/1750481317714121

Külling, C., Waller, G., Suter, L., Willemse, I., Bernath, J., Skirgaila, P., Streule, P., & Süss, D. (2022). *JAMES – Jugend, Aktivitäten, Medien – Erhebung Schweiz*. Zürich: Zürcher Hochschule für Angewandte Wissenschaften.

Ladeur, K.-H., & Gostomzyk, T. (2017). Gutachten zur Verfassungsmäßigkeit des Entwurfs eines Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz –NetzDG) i. d. F. vom 16. Mai 2017 – BT-Drs. 18/12356. https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf [22.04.2025].

Latzer, M., Just, N., Saurwein, F., & Slominski, P. (2003). Regulation remixed: Institutional change through self and co-regulation in the mediamatics sector. *Communications & Strategies, 50*(2), 127–157.

Lillian, D. L. (2007). A thorn by any other name: Sexist discourse as hate speech. *Discourse & Society, 18*(6), 719–740. https://doi.org/10.1177/0957926507082193

Marquart, F. (2023). Video killed the Instagram star: The future of political communication is audio-visual. *Journal of Visual Political Communication, 10*(1), 49–57.

Mayntz, R. (2004). Governance im modernen Staat. In A. Benz (Ed.), *Governance — Regieren in komplexen Regelsystemen: Eine Einführung* (S. 65–76). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90171-8_4

Merrill, S., & Åkerlund, M. (2018). Standing up for Sweden? The racist discourses, architectures, and affordances of an anti-immigration Facebook group. *Journal of Computer-Mediated Communication, 23*(6), 332–353. https://doi.org/10.1093/jcmc/zmy018

Müller, S. (2024, 25. Januar). GamerGate 2.0: Wie Streamer mit Hass Geld verdienen. *netzpolitik.org*. https://netzpolitik.org/2024/gamergate-2-0-wie-streamer-mit-hass-geld-verdienen/ [22.04.2025].

Müller, K., & Schwarz, C. (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. Social Science Research Network. https://ssrn.com/abstract=3149103

Müller, K., & Schwarz, C. (2018). *Making America hate again? Twitter and hate crime under Trump*. Social Science Research Network. https://ssrn.com/abstract=3149103 [22.04.2025].

Murthy, D., & Sharma, S. (2019). Visualizing YouTube's comment space: Online hostility as a networked phenomenon. *New Media & Society, 21*(1), 191–213. https://doi.org/10.1177/1461444818792393

Oehmer-Pedrazzi, F., & Pedrazzi, S. (2024). „An image hurts more than 1000 words?" Sources, channels, and characteristics of digital hate images. *Communications*, *49*(3), 421–443.

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication, 9*(1), 171–180.

Rabab'ah, G., Hussein, A. & Jarbou, S. (2024). Hate Speech in Political Discourse. *International Journal for the Semiotics of Law – Revue Internationale de Sémiotique Juridique, 37*, 2237–2256. https://doi.org/10.1007/s11196-024-10158-8

Rensinghoff, J. (2022). Die rechtliche Regulierung von Hass im Netz – Konzeption der Ehrverletzungsdelikte und ihr Schutz durch das NetzDG. In *Hate speech: Definitionen, Ausprägungen, Lösungen* (S. 275–292). Springer Fachmedien.

Romero-Rodríguez, L. M., Castillo-Abdul, B., & Cuesta-Valiño, P. (2023). The process of the transfer of hate speech to demonization and social polarization. *Politics and Governance*, *11*(2), 109–113.

Ruttloff, D., Haak, J., Groos, L., Moch, M., Mittler, N., Tophoven-Sedrakyan, T., & Borucki, I. (2022). Desinformation, Hassrede und Fake News – Wie viel Negativität verbreiteten die Parteien im Wahlkampf auf Social Media? In K.-R. Korte, M. Schiffers, A. von Schuckmann, & S. Plümer (Hrsg.), *Die Bundestagswahl 2021: Analysen der Wahl-, Parteien-, Kommunikations- und Regierungsforschung* (S. 1–42). Springer VS.

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, 146144482210911. https://doi.org/10.1177/14614448221091185

Schmid, U. K. (2023). Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*. https://doi.org/10.1177/14614448231198169

Schneiders, P. (2021). Hate speech auf Online-Plattformen: Problematisierung, Regulierung und Bewertung vor dem Hintergrund des Vorschlags für einen Digital Services Act. *UFITA, 86*(2), 266–327. https://doi.org/10.5771/2568-9185-2021-2-269

Schünemann, W. J., & Steiger, S. (2023). Die Regulierung von Internetinhalten am Beispiel Hassrede: Ein Forschungsüberblick. In S. Jaki & S. Steiger (Eds.), *Digitale Hate Speech: Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation* (S. 155–195). J. B. Metzler.

Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society, 21*(11), 1700–1714. https://doi.org/10.1080/1369118X.2017.1348535

Sponholz, L. (2018). *Hate speech in den Massenmedien*. Springer Fachmedien. https://doi.org/10.1007/978-3-658-15077-8

Stahel, L., Weingartner, S., Lobinger, K., & Baier, D. (2022). *Digitale Hassrede in der Schweiz: Ausmass und sozialstrukturelle Einflussfaktoren*. Universität Zürich. https://doi.org/10.21256/zhaw-26867

Udris, L., Rivière, M., Vogler, D. & Eisenegger, M. (2024). *Reuters Institute Digital News Report 2023. Länderbericht Schweiz*. Zürich: Forschungszentrum Öffentlichkeit und Gesellschaft (fög).

Unger, D., Unger-Sirsch, J. (2023). „Ihr gehört nicht dazu!" Soziale Ausgrenzung durch Hate Speech als Problem für liberale Demokratien. In S. Jaki & S. Steiger (eds), *Digitale Hate Speech*. J. B. Metzler. https://doi.org/10.1007/978-3-662-65964-9_9

Vergani, M., Martinez Arranz, A., Scrivens, R., & Orellana, L. (2022). Hate speech in a Telegram conspiracy channel during the first year of the COVID-19 Pandemic. *Social Media + Society, 8*(4). https://doi.org/10.1177/20563051221138758

Vogel, H. & Schmitt, E. (2024, 8. September). *TikTok als „scharfes Schwert" im Wahlkampf? Wieso die Plattform immer mehr in den politischen Fokus rückt*. swr.online. https://www.swr.de/swraktuell/baden-wuerttemberg/tiktok-soziale-medien-wahlkampf-afd-100.html

Walser, S. (2024, 17. Januar). Plattform X wird bedeutungslos. *Schweizer Radio und Fernsehen (SRF)*. https://www.srf.ch/news/wirtschaft/immer-mehr-verlassen-x-plattform-x-wird-bedeutungslos [22.04.2025].

Wilson, R. A. & Land, M. K. (2020). Hate speech on social media: Content moderation in context. *52 Connecticut Law Review 1029 (2021)*. https://ssrn.com/abstract=3690616 [22.04.2025].

Wirz, D. S, & Blassnig, S. (2024). Digitale Hassrede in der Schweiz. Eine Mehrmethodenstudie zur subjektiven und objektiven Konfrontation mit Hassrede im Alltag Schweizer Internetnutzer:innen. Abschlussbericht zu Händen des Bundesamts für Kommunikation. https://www.bakom.admin.ch/dam/bakom/de/dokumente/bakom/elektronische_medien/Zahlen%20und%20Fakten/Studien/digitale-hassrede-in-der-schweiz-bericht.pdf.download.pdf/Wirz-Blassnig%20(2024)%20Digitale%20Hassrede%20in%20der%20Schweiz.pdf [22.04.2025].

Yang, Z., Grenon-Godbout, N., & Rabbany, R. (2024). Game on, hate off: A study of toxicity in online multiplayer environments. *ACM Games, 2*(2), 1–13. https://doi.org/10.1145/3675805

Yildirim, M. M., Nagler, J., Bonneau, R., & Tucker, J. A. (2023). Short of suspension: How suspension warnings can reduce hate speech on Twitter. *Perspectives on Politics, 21*(2), 651–663. https://doi.org/10.1017/S1537592721002589