Ein Bild verletzt mehr als 1000 Worte?

Merkmale und Governance von Hassbildern im Netz.

Bericht zur Studie
November 2024

zu Händen des
Bundesamt für
Kommunikation BAKOM
Abteilung Medien
Dr. Thomas Häussler

von:

Franziska Oehmer-Pedrazzi | Fachhochschule Graubünden | franziska.oehmer@fhgr.ch
Stefano Pedrazzi | Université de Fribourg | stefano.pedrazzi@unifr.ch

Zitierempfehlung:

Oehmer-Pedrazzi, F. & Pedrazzi S. (2024). Ein Bild verletzt mehr als 1000 Worte? Merkmale und Governance von Hassbildern im Netz. Bericht zu Händen des Bundesamts für Kommunikation.

Gliederung

Warnhinweis: In dem Bericht werden Bilder oder Bildunterschriften zitiert und angezeigt, die Hass enthalten und somit starke negative Emotionen hervorrufen können. Um sich diesen Inhalten nicht aussetzen zu müssen, verweisen wir auf die Möglichkeit, die zentralen Ergebnisse des Berichts in der zusammenfassenden Darstellung (Executive Summary) zu rezipieren.

Executive Summary

Visuelle Inhalte wie Bilder, Memes oder Fotos stellen zentrale Ausdrucksformen der digitalen Kommunikation dar. Sie dienen nicht nur der Vermittlung von Informationen und der Unterhaltung, sondern können auch zur Verbreitung von Hassbotschaften genutzt werden. Unter dem Begriff *visuelle Hassrede* oder *Hassbilder* werden visuelle Ausdrucksformen (z. B. Fotos, Grafiken, Memes, Karikaturen) zusammengefasst, die dazu verwendet werden, Personen oder Gruppen aufgrund spezifischer Merkmale auszuschliessen, zu beleidigen, gegen sie zu Gewalt aufzurufen oder Gewalt gegen sie zu glorifizieren.

Der vorliegende Bericht widmet sich zwei zentralen Fragestellungen:

1. Welche charakteristischen Merkmale kennzeichnen visuelle Hassinhalte? (Teil I)
2. Welche Governance-Massnahmen sind geeignet, um visuellen Hass effektiv zu bekämpfen? (Teil II)

Teil I: Merkmale der Hassbilder

Der erste Teil des Berichts untersucht die Merkmale von Hassbildern im Netz. Im Fokus stehen dabei Intensitätsstufen des Hasses, Absender, internationale Bezüge, Hassobjekte und genutzte Kanäle in der visuellen Hasskommunikation in der Schweiz. Die Datengrundlage bilden online gefundene Hassbilder, die im Rahmen eines Citizen-Science-Ansatzes von der Schweizer Bevölkerung gesammelt und dem Forschungsprojekt bereitgestellt wurden. Die Analyse der Merkmale erfolgte durch eine standardisierte manuelle Inhaltsanalyse.

Ergebnisse

- Kanäle: Hassbilder werden nicht nur in sozialen Medien wie X (ehemals Twitter, 27,1 %) und Instagram (24,3 %) verbreitet, sondern auch auf bisher wenig untersuchten

Plattformen wie Amazon oder tutti.ch (7 %). Diese nutzerbasierten Plattformen haben hohe Reichweiten, bieten jedoch kaum Möglichkeiten für Nutzende, problematische Inhalte zu melden. Besorgniserregend ist zudem, dass auch publizistische Medien (10 %) durch die Abbildung von Hassbildern deren Reichweite erhöhen – selbst wenn die begleitenden Artikel kritisch sind.

- Adressaten/Ziele: Die Analyse zeigt, dass sich der Hass in den Bildern häufig gegen Personen aufgrund ihrer Nationalität (25 %) richtet. Auch die Geschlechtszugehörigkeit (21 %), insbesondere Transgenderpersonen (10,5 %), wird häufig angegriffen. Einstellungen zu Themen wie dem Ukraine-Krieg, Klimawandel oder Impfen dienen ebenfalls als Ziel von Anfeindungen.
- Intensität: Etwa die Hälfte der Hassbilder (50,7 %) nutzt sachliche oder humoristische Mittel, um andere auszugrenzen. Die andere Hälfte ist aggressiver, wobei 14 % strafrechtlich relevante Inhalte wie Tötungsaufrufe enthalten.
- Absender/Quellen: Hassbilder werden gleichermassen von ressourcenstarken Organisationen, darunter Parteien, und von Einzelpersonen verbreitet. Besonders problematisch ist die Beteiligung demokratischer Akteure wie Parteien und Politiker:innen an der Verbreitung solcher Inhalte.
- Internationalisierung: Viele der in der Schweiz verbreiteten Hassbilder greifen Themen aus Deutschland oder den USA auf oder haben ihren Ursprung in diesen Ländern. Dies verdeutlicht die internationale Dimension der Problematik.

Implikationen für die Governance:

Auf der Basis der Studienergebnisse können folgende Handlungsempfehlungen zur Governance von Hassbildern im Netz formuliert werden:

- Erweiterung der Regulierung: Die Studie zeigt, dass staatliche Massnahmen, die sich ausschliesslich auf grosse Kommunikationsplattformen wie X oder Instagram konzentrieren, nicht ausreichen. Plattformen mit geringerer Nutzung, aber mit bedeutendem Einfluss auf nationalem Level, wie beispielsweise Kleinanzeigenportale (tutti.ch), werden bisher vernachlässigt.
- Ausweitung rechtlicher Rahmenbedingungen: Hassbotschaften, die sich gegen Personen aufgrund spezifischer Ansichten oder Einstellungen richten, können nicht unter Bezugnahme von Artikel 261 des Schweizer Strafgesetzbuchs sanktioniert werden. Die gesetzliche Definition von Hassrede sollte erweitert werden, um auch

Inhalte einbeziehen zu können, die sich gegen spezifische (politische) Ansichten oder Einstellungen richten.

- Journalistische Verantwortung: Zusätzlich wird deutlich, dass publizistische Medien eine bedeutende Rolle bei der Verbreitung von Hassbildern spielen. Obwohl sie diese im Rahmen ihrer Berichterstattung oft kritisieren, tragen sie mit ihrer Reichweite zur weiteren Verbreitung bei. Eine verstärkte Selbstreflexion in den Redaktionen sowie Sensibilisierungsmassnahmen durch Branchenverbände sind daher anzuraten.
- Standards für politische Akteur:innen: Des weiteren zeigt die Studie, dass politische Akteure wie Parteien und Politiker:innen Hassbilder nutzen, um politische Gegner zu diskreditieren. Hier sollte über die Einführung von Standards für die Parteienkommunikation nachgedacht werden, ähnlich den Regelungen einiger Schweizer Parteien für die Verwendung von KI im Wahlkampf.
- Herausforderungen für Plattform-Selbstregulierung: Die Content Moderation von Hassbildern erfordert kontextuelles Wissen, das kultur- und zeitabhängig ist. Selbst mit den aktuellen Fortschritten in der Bilderkennung durch künstliche Intelligenz reichen technische und algorithmische Ansätze allein nicht aus, um visuelle Inhalte effektiv zu moderieren. Plattformen sollten für die manuelle Content Moderation auf Teams zurückgreifen, die divers aufgestellt sind (verschiedene Altersgruppen, kulturelle Hintergründe).
- Internationale Dimension von Hassbildern: Die internationale Verbreitung von Hassbildern zeigt, dass Governance-Massnahmen transnationale Kooperation erfordern, um die grenzüberschreitende Natur von digital vermitteltem (visuellem) Hass zu adressieren.

Teil II: Wirkung von Governance-Massnahmen

Im *zweiten Teil* des Berichts wird der Frage nachgegangen, welche *Governance-Massnahmen gegen Hassbilder* ergriffen werden können und welche Wirkungen sich damit erzielen lassen. Die Frage wird am Beispiel von Hassbildern, die sich gegen Transgender-Personen richten, im Rahmen einer Befragung im Experimentaldesign beantwortet. Im Fokus stehen dabei die Wirkungen von verschiedenen nutzendenzentrierten Governance-Massnahmen auf das Liken, Teilen, Weiterleiten, positive oder negative Kommentieren sowie Melden von Hassbildern. Zu den

untersuchten Governance-Massnahmen zählen: generische Counterbilder, die Hass allgemein verurteilen; spezifische Counterbilder, die inhaltliche und formelle Spezifika eines Hassbildes aufgreifen und gezielt kontern; empathische textliche Gegenrede sowie die Kennzeichnung von Hassbildern mittels eines «offiziellen» Warnhinweises (Labeling). Zudem wird unterschiedlichen Intensitätsstufen von visuell vermitteltem Hass Rechnung getragen.

Ergebnisse

- Wirkung von Governance-Massnahmen gegen Hassbilder: Die untersuchten Governance-Massnahmen können dazu beitragen, die Verbreitung von Hassbildern in sozialen Netzwerken und deren Weiterleitung über externe Kanäle zu begrenzen. Über alle untersuchten Governance-Massnahmen hinweg sind die Effekte jedoch moderat.
- Unterschiede in der Wirkung der Governance-Massnahmen: Empathische textliche Gegenrede sowie Counterbilder, die inhaltliche und gestalterische Spezifika eines Hassbildes aufgreifen, sind am wirksamsten hinsichtlich der Reduktion von Nutzendeninteraktionen wie Liken, Teilen, externes Weiterleiten oder positives Kommentieren. Generische Counterbilder und Warnhinweise durch Plattformen sind hingegen weniger wirksam. In Bezug auf öffentlich ablehnende oder die Verbreitung reduzierende Nutzendeninteraktionen wie negatives Kommentieren oder Melden von Hassbildern erweisen sich die untersuchten Governance-Massnahmen als nicht wirksam.
- Berücksichtigung der Intensität des Hassbildes: Die Intensität eines Hassbildes beeinflusst sowohl Nutzendeninteraktionen als auch die Wirksamkeit der Governance-Massnahmen. Humorvolle Hassbilder werden generell wahrscheinlicher geliked, geteilt, weitergeleitet und weniger wahrscheinlich gemeldet als aggressive oder Gewalt verherrlichende Hassbilder. Empathische textliche Gegenrede ist universell über alle Intensitätsstufen und positiv-verstärkenden Nutzendeninteraktionen hinweg wirksam und besonders effektiv bei Gewalt verherrlichenden Bildern. Spezifische Counter-Hassbilder zeigen sich als wirksamste Massnahme zur Reduktion verstärkender Nutzendeninteraktionen bei humorvollen und aggressiven Hassbildern, sie sind allerdings bei Gewalt verherrlichenden Bildern weniger effektiv. Generische Counter-Hassbilder erweissen sich am wenigsten wirksam, insbesondere bei

aggressiven und Gewalt verherrlichenden Bildern. Weitere Forschung ist nötig, um spezifische Effekte und widersprüchliche Ergebnisse besser zu verstehen.

Implikationen für die Governance Teil II:

Auf der Basis der Studienergebnisse können folgende Handlungsempfehlungen zur Governance von Hassbildern im Netz formuliert werden:

- Technische Massnahmen von Plattformen: Plattformen sollten nutzendenfreundliche Funktionen für visuelle und textbasierte Kommentierungen bereitstellen. Warnhinweise („Labeling“) sollten sparsam eingesetzt und auf gravierende Formen visueller Hassdarstellung wie Gewaltaufrufe beschränkt werden.
- Förderung der Kompetenzen von Nutzenden: Nutzende können die Verbreitung von Hassbildern eindämmen, indem sie gezielt mit Gegenrede und Counter-Hassbildern reagieren. Deshalb sollten sie durch Bildungsmassnahmen über die Auswirkungen ihrer Interaktionen aufgeklärt werden, um sie zu ermutigen, aktiv gegen Hassbilder vorzugehen.
- Rolle unabhängiger Organisationen: An «Trusted Flaggers» angelehnte «Trusted Commentators» könnten als neutrale Akteur:innen gezielt mit Gegenrede und Counter-Hassbildern eingreifen. Eine Finanzierung könnte durch öffentliche und private Mittel sowie durch Plattformen erfolgen.

1 Einleitung & Zielstellung

Der Stellenwert visueller Kommunikation nimmt seit vielen Jahren zu. Bilder gelten als zentrale Ausdrucksform digitaler Kommunikation (Hornuff, 2020, S. 16). Dies lässt sich zum einen auf den Bedeutungszuwachs von Plattformen und sozialen Medien wie Instagram oder YouTube, die die Verbreitung visueller Inhalte befördern, zurückführen (Marquart, 2023). Zum anderen 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). Zudem führen sie auch im Vergleich zu reinen textbasierten Beiträgen zu mehr Interaktionen (Schmid et al., 2022). So lässt sich bspw. auch das zahlreiche Teilen, Liken, Sharen und auch Modifizieren von sogenannten “Memes” – verstanden als Kombination von visuellen Inhalten, kurzen Texten oder Tags (vgl. Paciello et al., 2021) – erklären (Crawford et al., 2021). Zudem wird insbesondere Fotos ein hohes Mass an Authentizität zugesprochen (Graber & Lindemann, 2018, S. 62-63; auch Frischlich, 2018, S. 140). Mittels visueller Inhalte werden jedoch nicht nur wünschenswerte informative oder unterhaltende Inhalte kommuniziert. Sie dienen auch der bildlichen Vermittlung von Hassbotschaften, der Ausgrenzung von gesellschaftlichen Gruppierungen oder der Diffamierung Einzelner.

Der vorliegende Forschungsbericht widmet sich der Analyse solcher visueller Hassbotschaften: Im Zentrum stehen *zum einen* die Fragen nach den Merkmalen von Hassbildern, die online zirkulieren. Von Interesse ist dabei, welche Absender:innen, Hassobjekte, Intensitätsstufen, Stilmittel und internationale Bezüge sich in der visuellen Hasskommunikation in der Schweiz identifizieren lassen. Die Hassbilder wurden mittels eines Citizen-Science-Ansatzes gesammelt, bei dem die Schweizer Bevölkerung aufgerufen wurde, dem Forschungsprojekt gefundene online Hassbilder zur Verfügung zu stellen. Die Merkmale der Hassbilder wurden dann mithilfe einer standardisierten manuellen Inhaltsanalyse gewonnen. *Zum anderen* wird im zweiten Teil des Berichts auch der Frage nachgegangen werden, welche nutzendenzentrierten Governance-Massnahmen gegen Hassbilder ergriffen werden können und welche Wirkung sich damit erzielen lässt. Der Effekt verschiedener Massnahmen wurde im Rahmen einer Befragung im Experimentaldesign identifiziert. Auf der Basis der gewonnenen Erkenntnisse wurden jeweils Implikationen für die Governance von Hassbildern diskutiert.

2 Zum Phänomen Hass und Hassrede

Im akademischen und gesellschaftlichen Diskurs finden sich zahlreiche Definitionen des Begriffs „Hassrede“ (vgl. Brown 2017, S. 422; Sponholz, 2020). Einige beschränken den Begriff ausschliesslich auf rassistische Inhalte (Vaught, 2012) oder Aussagen, die sich gegen die ethnische oder religiöse Herkunft einer Person richten (Blaya et al., 2020). Andere Studien beziehen sich jedoch auch auf Einstellungen oder Positionen als Gruppenmerkmale, die angegriffen werden können (Räsänen et al., 2016). Der Begriff der digitalen Hassrede weist starke Parallelen zum Begriff „Cybermobbing“ auf. Obwohl sich beide Konzepte überschneiden, können sie im Allgemeinen dadurch unterschieden werden, dass Hassrede eine Form der Diskriminierung gegenüber einer Gruppe oder einer Person aufgrund ihrer Gruppenzugehörigkeit ist, während Cybermobbing eine wiederholte Angriffsform darstellt, die sich spezifisch gegen eine einzelne Person richtet (Räsänen et al., 2016).

Dieser Bericht definiert Hassrede in Anlehnung an Castaño-Pulgarín et al. (2021, S. 1) als „jede Kommunikation, die eine Person oder Gruppe aufgrund von Merkmalen wie Rasse, Hautfarbe, Ethnie, Geschlecht, sexueller Orientierung, Nationalität, Religion oder politischer Zugehörigkeit abwertet [eigene Übersetzung].“ Dementsprechend werden visuelle Hassrede oder Hassbilder als alle visuellen Ausdrucksformen (Fotos, Grafiken, Memes, Karikaturen usw.) verstanden, die andere aufgrund gruppenspezifischer Merkmale ausschliessen, beleidigen, zur Gewalt aufrufen oder Gewalt gegen andere Menschen verherrlichen.

Dass es sich dabei um kein Einzelphänomen handelt, zeigen Studien: In der Europäischen Union gaben 80 Prozent der Befragten an, dass sie online auf Hassrede gestossen sind. 40 Prozent haben sich selbst angegriffen und bedroht gefühlt (Gagliardone et al., 2015). Für die Schweiz zeigen Befragungsdaten, dass fast jede:r zehnte Befragte Anfeindungen im Netz ausgesetzt war (Stahel et al., 2022). Zwischen sieben und 23.4 Prozent der Kinder und Jugendlichen sind, so zeigt eine Metastudie, je nach Studiendesign, Land und Zeitraum Opfer von Hassrede (Kansok-Dusche et al., 2023).

Hass kann 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 Leiden verbunden (Stahel et al., 2022, S. 5). Im

Hinblick auf Jugendliche beobachteten Näsi et al. (2015) einen Verlust an sozialem Vertrauen, und Bilewicz & Soral (2020) stellten eine zunehmende Radikalisierung fest. Auf gesellschaftlicher Makroebene können Hassbotschaften ein Klima der Intoleranz und Angst schaffen (Stahel et al., 2022). 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).

3 Teil I: Merkmale von Hassbildern

3.1 Forschungsstand und Subfragestellungen zu den Merkmalen visueller «Hassrede»

Erkenntnisse zum Einsatz und den Merkmalen von visuellem Hass lassen sich bisher vor allem Studien entnehmen, die unterhaltsame oder sogar humorvolle Memes analysieren (Askanius, 2021; Sakki & Castrén, 2022; Schmid, 2023; Schmitt et al., 2020; Udupa, 2019). Demzufolge werden Memes v.a. auch von extremen politischen Organisationen eingesetzt, um ihre jeweiligen Ideologien zu verbreiten (Askanius, 2021). Ihr humoristischer Gehalt und die hohe Relevanz von Memes in der Alltagskommunikation würden die Identifikation als Hasskommunikation erschweren und die Hemmschwelle für das Teilen solcher Inhalte senken. In der Folge befürchtet man, dass dies zu einer Normalisierung extremistischer Ideen führen könne (Schmid, 2023). Weniger ist über andere visuelle Formate der Hassrede bekannt, die nicht explizit Memes sind und entweder in rein visueller Form oder in Kombination mit Text wirken, weshalb diese Studie auch diese Formen in den ersten beiden Forschungsfragen (F) einbezieht:

F1: Welche Formen visueller Hassrede lassen sich in der Schweiz finden?

F2: Wie lässt sich die Bild-Text-Interaktion innerhalb der gefundenen Formen visueller Hassrede beschreiben?

Die Literatur unterscheidet zwischen verschiedenen Intensitätsstufen von Hassrede: So kann zwischen harter, offener und sanfter, verdeckter Hassrede, legaler und illegaler Hassrede (Baider et al., 2020) oder Hassrede, die (international) strafrechtlich verboten und sanktioniert werden muss, oder als legale, aber intolerante Form der Meinungsäusserung auftritt, differenziert werden. In der Schweiz verbietet insbesondere

Artikel 261bis des Schweizerischen Strafgesetzbuchs öffentliche Handlungen oder Aussagen, die Menschen aufgrund ihrer Rasse, Ethnie, Religion oder sexuellen Orientierung herabsetzen oder diskriminieren. Die Diskriminierung von Personen aufgrund ihrer Einstellungen wird hingegen nicht darunter gefasst. In dieser Studie interessiert uns, welche Intensitätsstufen bei der Verbreitung visueller Hassrede in der Schweiz vorherrschen. Die dritte Forschungsfrage lautet daher:

F3: Welche Intensitätsstufen sind bei der Verbreitung visueller Hassrede in der Schweiz zu beobachten?

Bisherige Forschung zeigt, dass Hassrede oft aus strategischen Motiven meist von einflussreichen Akteuren verbreitet wird: Zu den Zielen zählen politischer Erfolg oder ökonomischer Gewinn. Sie kann durch koordinierte Netzwerke oder individuell verbreitet werden (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 beispielsweise 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 und mangelnde Empathiefähigkeit mit einer verstärkten Verbreitung von Hassrede verbunden (Frischlich et al., 2023). Daher zielt unsere vierte Frage darauf ab, die Quellen der visuellen Hassrede zu untersuchen:

F4: Wer sind die Quellen der Verbreitung visueller Hassrede in der Schweiz?

Die Forschung zu kleinen Mediensystemen wie der Schweiz zeigt, dass die Medienangebote von sogenannten „next door giants“ stark genutzt werden (Künzler, 2013) und dass entsprechende Spillover-Effekte beobachtet werden können, die auch für die Reezeption von Inhalten gelten. Besonders im Zusammenhang mit globalen Kommunikationsplattformen ist zu erwarten, dass auch bei der Verbreitung von Hassrede, die auf Vorurteilen und politischen Ideologien beruht und grenzüberschreitend

geteilt wird, Internationalisierungseffekte zu beobachten sind. Daraus ergibt sich unsere fünfte Forschungsfrage:

F5: Aus welchen geografischen Kontexten stammt die in der Schweiz verbreitete visuelle Hassrede?

Forschungsergebnisse zu den Zielgruppen von Hassrede zeigen, dass vor allem Mitglieder marginalisierter Gruppen von Online-Hassrede betroffen sind. 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 (Lillian, 2007; Sobieraj, 2018) oder ihres Flüchtlingsstatus (Kreis, 2017; Merrill & Åkerlund, 2018) angegriffen. Studien, die sich auf Jugendliche konzentrieren, zeigen zudem, dass beide Geschlechter gleichermassen von Hass betroffen sind, jedoch übernehmen Jungen häufig die Rolle des Aggressors (Wachs & Wright, 2020). Diese Studie zielt daher auch darauf ab, die Zielgruppen visueller Hassrede zu identifizieren. Die sechste Forschungsfrage lautet entsprechend:

F6: Welche Merkmale haben die Gruppen, die in der Schweiz durch visuelle Hassrede angegriffen werden?

Studien zu Hassrede haben Hassbotschaften auf einer Vielzahl von Kanälen gefunden und analysiert. Die meisten Studien konzentrieren sich auf grosse Kommunikationsplattformen wie Twitter (Burnap & Williams, 2015), Facebook (Farkas et al., 2018; Merrill & Åkerlund, 2018) und YouTube (Murthy & Sharma, 2019), die für die Datenerhebung vergleichsweise zugänglicher sind (oder waren). Weniger häufig jedoch ist Forschung, die die Verbreitung von Hassbotschaften in rechten und extremen Online-Communities untersucht (Askanius, 2021; Rieger et al., 2021). Die Verbreitung von Hass über (private) Kommunikationskanäle, wie Messaging-Dienste (bspw. WhatsApp oder Telegram), ist noch weniger zugänglich, könnte jedoch eine besonders wichtige Rolle bei der Verbreitung visueller Hassrede spielen, da diese als privater Raum wahrgenommen werden, der als geschützt gilt (Vergani et al., 2022). Im Gegensatz dazu ist die Verbreitung von Hass durch journalistische Medien (Harlow, 2015) oder in Kommentarsektionen journalistischer Medien dokumentiert (Paasch-Colberg et al., 2021). Das Ziel dieser

Studie ist es, die Kanäle zu identifizieren, über die visuelle Hassrede verbreitet wird. Unsere siebte Forschungsfrage lautet:

F7: Über welche Kanäle wird visuelle Hassrede in der Schweiz verbreitet?

3.2 Methode und Design zur Analyse der Merkmale von Hassbildern

Zur Beantwortung der Forschungsfragen wurde eine manuelle, quantitative Inhaltsanalyse angewandt. Automatisierte Verfahren wurden aufgrund des hohen Interpretationsaufwands, den visuelle Inhalte im Vergleich zu Texten erfordern, als weniger geeignet eingestuft (Schwertberger & Rieger, 2021).

3.2.1 Der Prozess der Datenspende

Die Hassbilder wurden im Rahmen eines Citizen-Science-Ansatzes gesammelt: Vom 3. Februar bis zum 3. März 2023 wurde eine einmonatige Kommunikationskampagne durchgeführt, bei der die Schweizer Bevölkerung aufgefordert wurde, digitale Hassbilder über die Website www.hassbilder-verletzen.ch einzureichen. Diese Informationskampagne wurde in Zusammenarbeit mit etablierten zivilgesellschaftlichen Organisationen durchgeführt, die in diesem Bereich tätig sind. Zu diesen Organisationen gehören unter anderem „alliance F“ [der grösste Schweizer Frauen-Dachverband], der Verband der Schweizer Jugendparlamente, das Museum für Kommunikation Bern, das Fotomuseum Winterthur oder auch die Stiftung gegen Rassismus und Antisemitismus.
Die Kampagne begann mit einer Pressekonferenz und einer Podiumsdiskussion, an der Expert:innen aus der Forschung, ein Vertreter einer Plattform und ein Diversity-Beauftragter teilnahmen. Aufgrund eines Hackerangriffs auf die Projektwebsite erregte das Forschungsprojekt sowohl in regionalen als auch in nationalen Medien Aufmerksamkeit. Zudem wurde die Kampagne auf sozialen Medien, vor allem Twitter (jetzt X), LinkedIn, Instagram und über Newsletter kooperierender zivilgesellschaftlicher Organisationen sowie der Studienautor:innen durchgeführt. In den Städten Bern und Zürich wurden zudem in einigen Quartieren Flyer verteilt. Weiterhin wurde ein Aufruf auf der Citizen-Science-Plattform „Schweiz forscht“ veröffentlicht.
Dieser bürgerwissenschaftliche Ansatz der Datensammlung hat den Vorteil, dass – anders als bei vielen anderen Studienansätzen – keine Beschränkung der Analyse auf

einen oder wenige Kanäle (wie Twitter, Facebook oder Telegram), auf spezifische Themen wie Migration (Paasch-Colberg et al., 2021) oder auf rassistische Inhalte (Hangartner et al., 2021) oder auf wenige Nutzerkonten notwendig ist.

Die Citizen Scientists wurden gebeten, die Hassbilder per Drag-and-Drop auf die Website hochzuladen und unmittelbar danach Fragen zum gespendeten Bild und zu ihrer eigenen Person zu beantworten. Die Website enthielt eine Definition von Hassrede und verwies darauf, dass sich diese gegen Merkmale einer Gruppe und nicht gegen Einzelpersonen ohne Bezug zur Gruppenzugehörigkeit richten muss. Ausserdem wurde ein Beispiel-Screenshot eines verpixelten Hassbildes angezeigt, um zu verdeutlichen, welche Informationen auf dem Bild erkennbar sein sollten, um dessen Analyse zu erleichtern. Dazu gehören Details wie der Name der Quelle, die Anzahl der Interaktionen oder das Veröffentlichungsdatum. Die Klarheit und Benutzerfreundlichkeit der Website wurde durch Pretests mit fünf Personen unterschiedlichen Bildungsstandes und Alters überprüft.

Um ein Verständnis der soziodemografischen Zusammensetzung der Citizen Scientists zu gewinnen, wurden die Teilnehmenden auch nach ihrem Alter, Geschlecht und Bildungsstand gefragt. Von den 72 Bildern, die über die Projektwebsite eingereicht wurden, liegen zu 50 Bildern Informationen über die Citizen Scientists vor. Das Durchschnittsalter betrug 35 Jahre, wobei der jüngste Datenspender 20 Jahre alt war und der älteste 65. Ältere Personen sind somit unter den Datenspendern deutlich unterrepräsentiert. Das Geschlechterverhältnis ist relativ ausgeglichen: 25 Frauen, 24 Männer und eine Person, die sich als „divers“ identifiziert, nahmen teil. Menschen mit einem Universitätsabschluss sind leicht überrepräsentiert, da mehr als die Hälfte (n=27) der Bilder von Personen mit akademischem Hintergrund eingereicht wurde. Vier Personen hatten eine Berufsausbildung und eine Person die obligatorische Schule abgeschlossen.

Als Ergebnis der Kampagne wurden insgesamt 86 Hassbilder gespendet. Davon wurden 15 im Rahmen eines Datenaustauschs mit der Plattform „Meldestelle für Rassismus im Internet“ eingereicht, die ebenfalls anonyme Spenden über ein Online-Formular ermöglicht.

3.2.2 Einstufung als Hassbild

Zunächst musste entschieden werden, ob das eingereichte Bild als Hassbild qualifiziert werden kann oder nicht. Aufgrund des hohen Interpretationsaufwands bei visuellen Inhalten ist diese Kategorisierung unvermeidlich subjektiv und kann auch durch kulturelle Faktoren beeinflusst werden. Um jedoch die höchstmögliche Intersubjektivität bei der Klassifizierung zu erreichen, wurde ein Kriterienkatalog mit Ja-oder-Nein-Entscheidungsfragen für die Codierer entwickelt. Je mehr Fragen mit „Ja“ beantwortet werden können, desto wahrscheinlicher wird das Bild als Hassbild wahrgenommen. Die Grundlage für den Kriterienkatalog ist der Fünf-Punkte-Test des Ethiknetzwerks für Journalisten, der ursprünglich entwickelt wurde, um Hassrede zu erkennen (https://ethicaljournalismnetwork.org/point-two). Die Anpassungen des ursprünglichen Fünf-Punkte-Tests bestanden hauptsächlich aus sprachlichen Änderungen oder inhaltlichen Kürzungen, um der wissenschaftlichen Zielsetzung gerecht zu werden. Die folgenden Fragen bildeten die Grundlage für die Klassifizierungsentscheidung (Ja oder Nein Hassbild):

- Wer ist die Quelle des Hassbildes (sofern sichtbar)? Lässt der Status oder Ruf der Quelle die Veröffentlichung von Hassbotschaften wahrscheinlich erscheinen?
- Handelt es sich um eine öffentliche Kommunikation, die sich an ein disperses Publikum richtet? (Anmerkung: Hassbotschaften erfordern Öffentlichkeit.)
- Ist die Absicht erkennbar, mit dem Bild Einzelpersonen aufgrund eines gruppenspezifischen Merkmals oder Gruppen zu schaden?
- Ist der Inhalt des Bildes geeignet, Hass gegen andere zu säen, andere auszugrenzen oder zu diskreditieren, oder sogar Gewalt zu provozieren?
- Ist der Ton der Sprache oder die Gestaltung des Bildes aggressiv, zum Beispiel durch Schimpfwörter oder Symbole?

Zusätzlich wurden die Antworten der Datenspender:innen genutzt. Falls erforderlich, wurde eine ergänzende Desktop-Recherche durchgeführt. Der hohe Reliabilitätswert (Krippendorff's $\alpha = 0{,}93$) bestätigt den Nutzen des gewählten Ansatzes.

3.2.3 Analysierte Variablen

Für die als Hassbilder klassifizierten Bilder wurden Merkmale in Bezug auf formale Eigenschaften (Veröffentlichungskanal, Form der visuellen Kommunikation) und Inhalt (Intensität des Hasses, Bild-Text-Interaktion), Quelle und Ziel des Hasses von zwei Codierern erfasst. Nachfolgend werden die formalen Variablen und ihre Kategorien in tabellarischer Form vorgestellt, während die eher inhaltlichen Merkmale der Inhaltsanalyse im Text detaillierter dargestellt werden.

Tabelle 1: Analysierte formale Variablen

Variable	Ausprägungen
Form der visuellen Kommunikation (α =0.93)	1 = Fotografie 2 = Meme: Dabei handelt es sich um eine Kombination von visuellen Inhalten, kurzen Texten oder Tags (vgl. Paciello et al., 2021) 3 = Karikatur: bildliche Form der Satire 4 = Grafik/Zeichnung / Plakat 5 = Grafik (nur Text) / Plakat 6= Symbol/Piktogramm 9 = Sonstiges
Kanal/Plattform Über welchen Kanal wurde das Hassbild verbreitet? (α =0.93)	Mit dieser Variable wird der Verbreitungskanal des Hassbildes erfasst. Als Grundlage dient die von der /vom Datenspender:in zur Verfügung gestellte Information. *Social Media* 1 Facebook 2 Instagram 3 Twitter 4 TikTok 9 Sonstige Social Media *Publizistisches Medium* 10 Kommentar/Leserforum eines Mediums 11 Publizistisches Medium 12 Alternatives Medium: Als alternative Medien werden solche Nachrichtenangebote verstanden, die sich selbst aktiv als Vertreter einer Gegenposition zu den «mainstream» Medien bezeichnen (Schwaiger, 2022). 19 Sonstiges Medium *Andere Onlinepublikationen* 20 Blog 21 Webseite 29 Sonstige Onlinepublikation *Messengerdienst* 30 WhatsApp 31 Telegram 39 Sonstiger Messengerdienst 99 nicht erkennbar

Die Variablen wurden deduktiv und induktiv aus selbst recherchierten Hassbildern aus anderen Ländern entwickelt.

- Bild-Text-Interaktion: Inhaltlich wurde erfasst, wie das Bild und der Text (falls vorhanden) miteinander interagieren (Bild-Text-Interaktion α = 0.87). Es wurde

unterschieden zwischen Bildern, bei denen der Text allein den Hass vermittelt, während das Bild nur zur Veranschaulichung dient; Bildern, die Hass ohne begleitenden Text vermitteln; Bildern, die sowohl visuell als auch textlich diskriminieren; und Hassbildern, bei denen der Hass durch das Zusammenspiel von Bild und Text vermittelt wird.

- Intensität des Hasses: Inhaltlich wurde auch die Intensität des Hasses (α = 0.87) bewertet, die das Bild zum Ausdruck bringt. Nicht-aggressive Diskriminierungen, wie zum Beispiel Karikaturen oder oft auch Memes, wurden von aggressiven Diskriminierungen unterschieden, bei denen Gruppen durch den Einsatz von Schimpfwörtern oder Vergleiche mit Tieren, Hitler oder dem Dritten Reich negativ dargestellt werden. Die höchste Intensitätsstufe wurde Bildern zugewiesen, die Gewalt verherrlichen oder zu Gewalt aufrufen, was möglicherweise rechtliche Fragen und Bedenken aufwirft.
- Quelle: Die Quelle wurde als individuelle:r Akteur:in mit oder ohne Angabe des realen Namens oder als kollektiver Akteur (Akteurtyp: α = 0.97) kategorisiert. Für die Analyse war auch von Interesse, ob die Quelle als mutmasslicher Ursprung des Bildes angesehen werden kann oder ob das Bild weitergeleitet wurde (Verbreitungsform: α = 0.83). Wenn Letzteres der Fall war, wurde vermerkt, ob die Quelle der Nachricht in dem Bild eine zustimmende oder ablehnende Haltung gegenüber dem Inhalt hatte (Haltung zum weitergeleiteten Bild: α = 0.90).
- Zusätzlich wurde, um Annahmen über die Internationalisierung von Hassbildern zu treffen und angesichts der Tatsache, dass die Schweiz von kulturell und sprachlich ähnlichen Ländern umgeben ist, das Herkunftsland der Quelle (α = 0.78) und die verwendete Sprache (α = 1.00) erfasst.
- Ziele des Hasses: Weiter wurde eine Variable codiert, um zu erfassen, gegen welche gruppenspezifischen Merkmale (α = 0.87) sich das Hassbild richtete. Basierend auf einem breiten Verständnis von Hassrede wurden Unterscheidungen getroffen zwischen Merkmalen, die in die Kategorien Körper & Gesundheit (Frauen, Männer, Transgender, sexuelle Orientierung, Alter, Hautfarbe, Gesundheit), Kultur & Gesellschaft (Religion, Nationalität, Beruf, wirtschaftlicher Status) und Positionen (Haltung zur Impfung, zum Ukraine-Konflikt, zum Klimawandel) fallen.

3.4 Ergebnisse: Merkmale der Hassbilder

Insgesamt wurden 86 Bilder als potenzielle Hassbilder eingereicht. Im Rahmen der Kodierung konnten jedoch neun dieser Bilder nicht als Hassbilder eingestuft werden und wurden daher von der Analyse ausgeschlossen. Darunter befanden sich beispielsweise Screenshots von Chatprotokollen oder dokumentarische Fotos von Hasshandlungen. In beiden Fällen handelte es sich nicht um visuelle Inhalte, die eigenständig dem Zweck dienten, Hass zu verbreiten. In einigen Fällen war für die Kodierung zusätzliche Hintergrundinformation notwendig. Die Daten wurden mit SPSS analysiert. Neben Prozentwerten und absoluten Zahlen wurden auch die minimalen und maximalen Werte der 95%-Konfidenzintervalle berichtet, um eine ungefähre Vorstellung von der Verteilung in der Bevölkerung zu geben. Aufgrund der relativ kleinen Stichprobengrösse sind die Intervalle teilweise gross.

3.4.1 Formen visuellen Hasses

Die erste Forschungsfrage beschäftigte sich mit den verschiedenen Erscheinungsformen visuellen Hasses, indem unterschiedliche Formen wie Fotografien, Memes, Grafiken, Karikaturen und andere unterschieden wurden. Bei der Mehrheit der eingereichten Bilder handelte es sich um Grafiken (38.2 %, n = 29, 95%-KI [27.6; 48.7]). Memes folgen dicht dahinter (35.5 %, n = 27, 95%-KI [25.0; 47.4]) und spiegeln die wichtige Rolle wider, die Memes in der aktuellen Online-Kommunikation spielen. Fotografien machen 15.8 % des analysierten Materials aus (n = 12, 95%-KI [7.9; 25.0]). Es gab nur vier Karikaturen in der Stichprobe (5.3 %, 95%-KI [1.3; 10.5]).

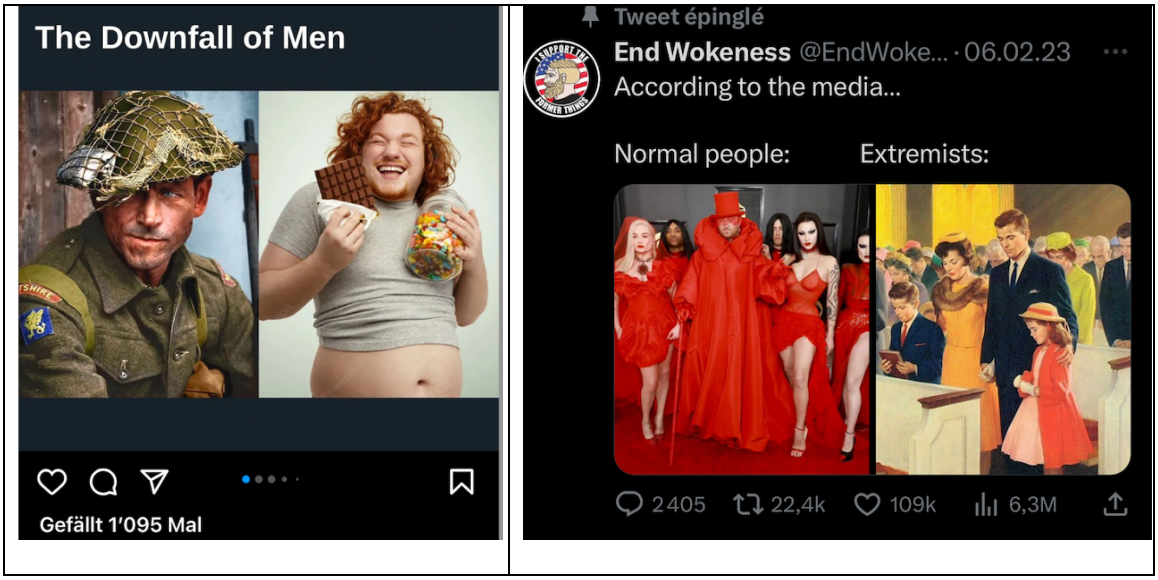
3.4.2 Ergebnisse zur Bild-Text-Interaktion

Es wurde auch untersucht, inwieweit die Bilder Textelemente enthielten und welche Rolle der Text bei der Vermittlung der Botschaft spielte. Von Interesse war dabei, ob der Text allein die Hassbotschaft vermittelte und das Bild nur als Illustration und Aufmerksamkeitserreger diente, ob das Bild allein Hass verbreitete, ob sowohl Bild als auch Text gleichermassen Hass vermittelten oder ob der Hass nur durch das Zusammenspiel von Bild und Text entstand.

Zunächst ist festzustellen, dass alle eingereichten Bilder auch Textelemente enthielten. In mehr als der Hälfte der Fälle (54.7 %, n = 41, 95%-KI [42.7; 65.3]) entsteht der Hass durch die Interaktion zwischen Text und Bild – das bedeutet, dass weder das Bild noch der Text für sich allein als Hass erkannt worden wären.

In fast jedem dritten Hassbild (30.7 %, n = 23, 95%-KI [20.0; 41.3]) dient der visuelle Inhalt nur zur Veranschaulichung. Er wird offenbar nur verwendet, um Aufmerksamkeit zu erregen. Neun Bilder vermitteln Hass sowohl visuell als auch textlich (12.0 %, 95%-KI [5.3; 20.0]). Nur zwei der eingereichten Bilder (2.7 %, 95%-KI [0; 6.7]) können Hass ausschliesslich durch visuelle Mittel vermitteln.

Abbildung 1: Beispiele für eine zusammenspielende Bild-Text-Interaktion



3.4.3 Ergebnisse zur Intensität des vermittelten Hasses

Die dritte Forschungsfrage trägt dem Umstand Rechnung, dass Hass in unterschiedlicher Intensität vermittelt werden kann. Diese kann in aufsteigender Intensität klassifiziert werden: von Hassbildern, die in einer nicht-aggressiven, d. h. sachlichen oder sogar humorvollen Weise andere Gruppen von Menschen ausschliessen, über aggressive Diskriminierung, d. h. die Ausgrenzung durch die Verwendung von Schimpfwörtern, entmenschlichenden Darstellungen oder Vergleichen, bis hin zu Bildern, die Gewalt

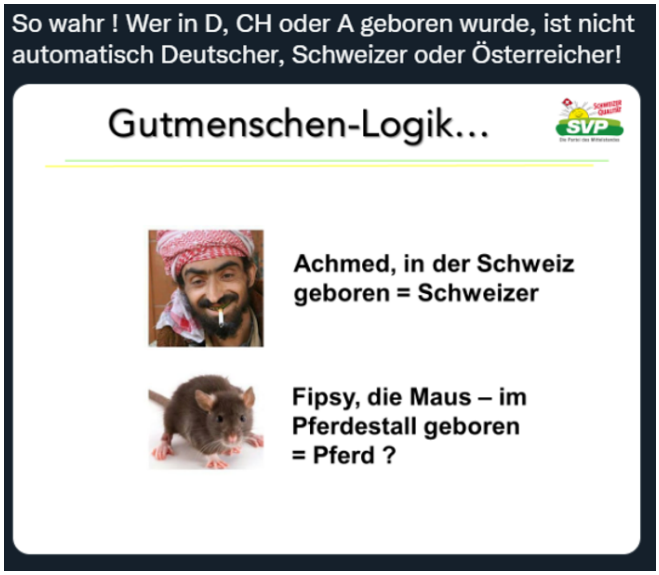
verherrlichen oder zu Gewalt aufrufen. Letztere können auch rechtliche Bedenken aufwerfen.

Bei der Hälfte der Bilder (50.7 %, n = 38; 95%-KI [40.0; 62.7]) handelt es sich um eine nicht-aggressive Form von Hassbildern. Hier werden andere Gruppen von Menschen auf sachliche oder humorvolle Weise ausgeschlossen oder beleidigt.

Abbildung 2: Beispiel für eine humorvolle Diskriminierung

Aggressive, also stark beleidigende und entwürdigende Hassbilder machen 34.7 % der Stichprobe aus (n = 26, 95%-KI [24.0; 45.3]). Dies umfasst zum Beispiel entmenschlichende Darstellungen von Personengruppen (siehe Abbildung 3). Ein Drittel der aggressiven diskriminierenden Hassbilder (n = 8) wurden von Organisationen wie politischen Parteien verbreitet.

Abbildung 3: Beispiele für aggressive Diskriminierung durch Entmenschlichung



In fünf eingereichten Hassbildern (6.7 %, 95%-KI [1.3; 13.3]) wird das (Selbst-)Töten oder der Mord an anderen Gruppen explizit goutiert. Zudem rufen drei Hassbilder (4.0 %, 95%-KI [0; 9.3]) – rechtlich strafbar – zum Töten auf.

Abbildung 4: Beispiel für die Verherrlichung von Gewalt

3.4.4 Ergebnisse zu den Quellen der Hassbilder

Zudem wurden auch die Quellen bzw. Urheber der visuellen Hassbilder untersucht: In 33.8 % der Fälle (n = 25, 95%-KI [23.0; 44.6]) werden Hassbilder aktiv von Organisationen und kollektiven Akteuren wie politischen Parteien oder Medienorganisationen verbreitet, häufig von Akteuren, die für Kommunikationsaufgaben über finanzielle und personelle Ressourcen verfügen. Unter den gespendeten Bildern befinden sich Hassbilder, die von politischen Parteien veröffentlicht wurden. Einzelpersonen verbreiten ebenfalls Hassbilder, sowohl mit als auch ohne Offenlegung ihres realen Namens, und zwar in gleichem Masse (21.6 %, n = 16, 95%-KI [12.2; 31.1]). Dies schliesst auch Politiker:innen ein, die Hassbilder von ihren persönlichen Accounts aus teilen. In mehr als der Hälfte der Fälle (55.3 %, n = 42, 95%-KI [43.4; 65.8]) stammen die Bilder auch von den mutmasslichen Ursprungsquellen. In jedem fünften Fall (21,1 %; n = 16, 95%-KI [11,8; 30,3]) werden die Bilder von einer anderen Quelle geteilt. Wenn die Bilder weitergeleitet wurden, wurde auch aufgezeichnet, ob dies in Zustimmung oder Ablehnung geschah: In acht Fällen (36.4 %, 95%-KI [18.2; 54.5]) war eine klare Zustimmung erkennbar, während in fünf Fällen (22.7 %, 95%-KI [9.1; 40.9]) das Hassbild kritisch kommentiert wurde.

3.4.5 Ergebnisse zur Internationalisierung von Hass

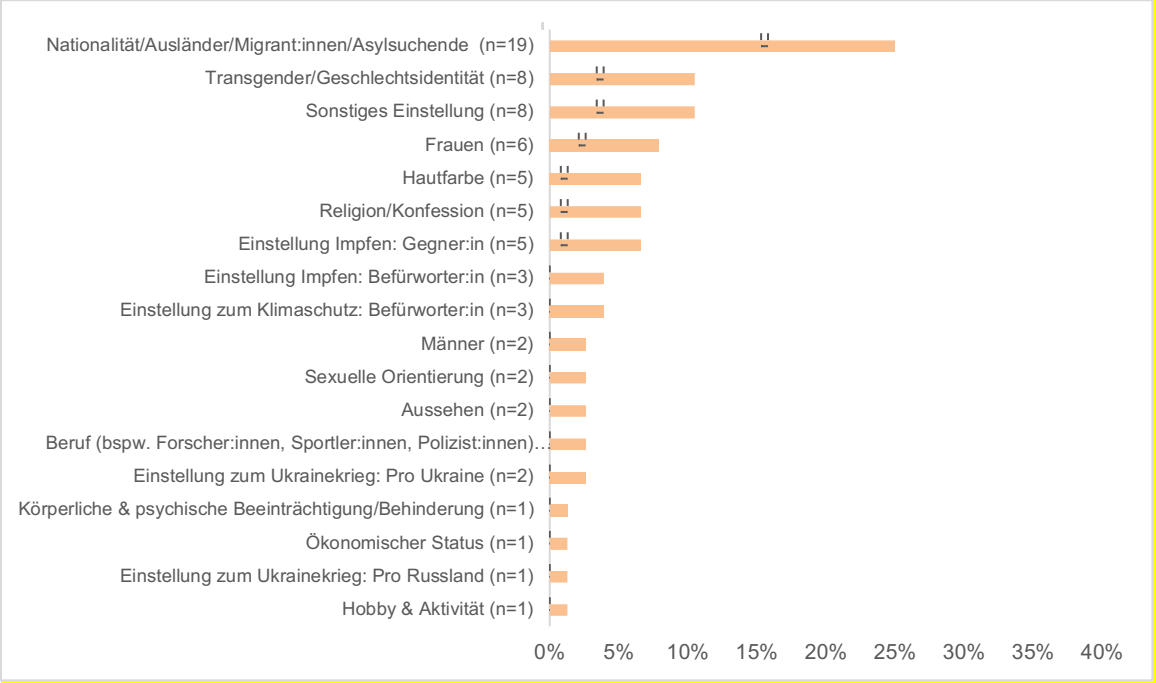
Im Kontext des Kleinstaates Schweiz war es auch von Interesse, den geografischen Kontext der Hassbilder zu bestimmen. Von besonderem Interesse ist, ob Einflüsse aus anderen Ländern oder Sprachregionen erkennbar sind. Wann immer möglich, wurden das Herkunftsland der Quellen oder der nationale Kontext der dargestellten Themen und Ereignisse aufgezeichnet. Visuelle Indikatoren für die Klassifizierung waren nationale Flaggen oder Symbole sowie nationale Figuren und Charaktere wie Heidi oder Globi (bekannte Schweizer Kinderbuchfiguren). Die Daten zeigen, dass die Verbreitung von Hass auch eine europäische oder sogar internationale Dimension hat. Besonders für ein kleines Land wie der Schweiz, das von grossen Nachbarländern umgeben ist, wurde deutlich, dass acht der 39 eingereichten Hassbilder aus Deutschland stammen (20.5 %, 95%-KI [7.7; 33.3]) und drei starke Verbindungen zu den USA aufweisen (7.7 %, 95%-KI [0; 17.9]).

Ein weiterer Indikator für die potenzielle Internationalisierung von Hassbildern ist die verwendete Sprache. Die Mehrheit der Texte in den eingereichten Hassbildern war auf Deutsch geschrieben (65.3 %, n = 49, 95%-KI [53.3; 74.7]). Nur eines der Bilder war in Schweizerdeutsch verfasst. Mehr als ein Viertel der Bilder verwendete die englische Sprache (28,0 %, n = 21, 95%-KI [18.7; 38.7]). Es wurden kaum Bilder in den anderen offiziellen Landessprachen der Schweiz, Französisch und Italienisch, eingereicht. Dies könnte jedoch auch auf den starken Kampagnenfokus auf die Datensammlung im deutschsprachigen Teil der Schweiz zurückzuführen sein.

3.4.6 Ergebnisse zu den Adressaten der Hassbilder

Von besonderem Interesse in dieser Studie war die Frage, welche Gruppen von Menschen durch Hassbilder angegriffen werden. Am häufigsten von Hassbildern betroffen sind Ausländer:innen und Migrant:innen. Jedes vierte eingereichte Hassbild richtete sich gegen Personen anderer Nationalitäten (25.0 %, n = 19, 95%-KI [15.8; 35.5]). 21 % der eingereichten Bilder zeigen Hass gegen Menschen aufgrund ihres Geschlechts oder ihrer Geschlechtsidentität. Besonders häufig sind transgeschlechtliche Personen Ziel von Angriffen (10.5 %, n = 8, 95%-KI [3.9; 17.1]). Frauen werden ebenfalls häufig angefeindet (7.9 %, n = 6, 95%-KI [2.6; 14.5]). Zwei Bilder zielten auf Männer ab (2.6 %, 95%-KI [0; 6.6]). Hautfarbe und Religion waren ebenfalls Gründe für visuelle Hassangriffe (jeweils 6.6 %, n = 5, 95%-KI [1.3; 13.2]). Verschiedene Positionen zu aktuellen Themen wie Klimaschutz, dem Ukraine-Konflikt oder der Impfung waren ebenfalls Motive für die Verbreitung von Hassbildern gegen die gegnerische Seite (siehe Abbildung 5).

Abbildung 5: Adressaten visuellen Hasses (n = 76)



Anmerkung: Die Grafik visualisiert, wie sich die eingereichten Hassbilder auf verschiedene Zielgruppen verteilen (in Prozent).
Lesebeispiel: 19 der 76 eingereichten Hassbilder, dies entspricht einem Anteil von 25%, richteten sich gegen Personen aufgrund ihrer Nationalität oder ihres Status' als Migrant:innen oder Asylsuchende.

3.4.7 Kanäle, über die Hassbilder verbreitet werden

Die am häufigsten verwendeten Kanäle zur Verbreitung der eingereichten Hassbilder waren Twitter/X (27.1 %, n = 19, 95%-KI [17.1; 37.1]) und Instagram (24.3 %, n = 17, 95%-KI [14.3; 34.3]). Diese 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.

Hassbilder finden auch in journalistischen Medien Verbreitung: Sieben von 70 eingereichten Hassbildern, bei denen der Kanal identifizierbar oder erwähnt wurde, wurden auf den Websites von Zeitungen, Fernsehsendern oder Online-Medien veröffentlicht (10.0 %, 95%-KI [2.9; 17.2]). Obwohl die Botschaften dieser Hassbilder oft in begleitenden Artikeln kritisch hinterfragt werden, ist ihre Veröffentlichung in den Medien doch mit einer zusätzlichen – nicht unerheblichen – Reichweite. Weitere Websites

wie Reddit, Amazon oder tutti.ch dienen ebenfalls als Plattformen für Hassbilder (7.0 %, n = 5, 95%-KI [1.4; 12.9]). Beispielsweise wurde auf dem Kleinanzeigenportal tutti.ch ein „N***“ zum Verschenken angeboten, mit der Beschreibung: „Der N*** ist praktisch. Am liebsten hat er es in der Gaskammer.“ Diese Plattformen, die in der Regel eine hohe Nutzendenzahl aufweisen, stehen nicht im primären Fokus der Kommunikationsforschung oder der Regulierungsmassnahmen politischer und zivilgesellschaftlicher Akteure. Nur je ein Hassbild wurde auf TikTok und WhatsApp verbreitet (1.4 %, 95%-KI [0; 4.3]).

3.5 Fazit und Implikationen für die Governance von Hassbildern

Der erste Teil des Berichts konzentrierte sich auf die Analyse der Merkmale visueller Hassbilder. Die Hassbilder wurden im Rahmen einer Kampagne gesammelt, bei der die Schweizer Bevölkerung eingeladen wurde, im Sinne von Citizen Science teilzunehmen. Anschliessend wurden die Bilder mittels standardisierter manueller Inhaltsanalyse untersucht. Mit wenigen Ausnahmen wurden alle eingereichten Bilder während des Kodierungsprozesses als Hassbilder bestätigt. Dies deutet darauf hin, dass Nutzende ein intuitives Gespür dafür haben, potenziell diskriminierende und schädliche Bilder zu erkennen, auch ohne formelle Schulung oder tiefere Kenntnisse des Themas.

Visueller Hass tritt in verschiedenen Formen auf, wobei Grafiken besonders häufig verwendet werden. Auch Memes werden zur Verbreitung von Hass genutzt (F1). Visueller Hass funktioniert überwiegend durch die Interaktion mit Text oder allein durch den Text. Reine visuelle Inhalte ohne Text sind kaum anzutreffen (F2). Etwa die Hälfte der Hassbilder nutzt subtile oder humorvolle Mittel, um andere Personen oder Gruppen auszuschliessen, was in der Regel als freie Meinungsäusserung gilt und nicht rechtlich verfolgt wird (F3). Ein Teil der Hassbilder nimmt hingegen einen aggressiven Ton ein, einige enthalten sogar explizit illegale Inhalte wie Aufrufe zur Gewalt oder zum Mord.

Hassbilder werden (F4) sowohl von vergleichsweise ressourcenstarken Organisationen wie politischen Parteien als auch von Einzelpersonen (auch unter Verwendung ihrer realen Namen) verbreitet. Besonders auffällig ist, dass politische Parteien als zentrale Akteure der Demokratie Hassbilder verbreiten. Die internationale Dimension von Hassbildern (F5) wurde deutlich, da sich Quellen und thematische Schwerpunkte häufig auf Deutschland oder die USA zurückführen lassen. Einige Hassbilder enthalten Text in

englischer Sprache. Hassobjekte (F6) sind vor allem ausländische sowie transgeschlechtliche Personen. Der hohe Anteil an ausländerfeindlichen Hassbildern könnte auch darauf zurückzuführen sein, dass eine Reihe von Hassbildern über die Plattform „Report Online Racism“ zur Verfügung gestellt wurde. Auch Vorurteile und politische Positionen der Menschen dienen als Gründe für Feindseligkeit und können als Indikatoren für eine polarisierte Gesellschaft interpretiert werden.

Die Ergebnisse beleuchten auch die Kanäle (F7), über die Hassbilder verbreitet werden: Neben etablierten Kommunikationsplattformen wie Twitter, Instagram und Facebook wurde deutlich, dass auch Plattformen, die bislang weder in der Forschung noch in der Regulierung besonders berücksichtigt wurden, als Katalysatoren für Hass dienen. Besonders problematisch ist die Rolle journalistischer Medien, die Hassbilder durch ihre weite Verbreitung zusätzlich verstärken, selbst wenn sie diese in begleitenden Artikeln kritisch diskutieren. Dies ist besonders problematisch, da Bilder oft mehr Aufmerksamkeit erregen und die kritischen Bemerkungen im Text möglicherweise weniger stark wahrgenommen und erinnert werden. Zudem kann die Darstellung von Hassbildern in traditionellen Nachrichtenmedien zu einer zunehmenden Normalisierung und Verbreitung diskriminierender Ideologien führen (Schmid, 2023; Schmitt et al., 2020). Des Weiteren wurden Hassbilder analysiert, die nicht auf Kommunikationsplattformen, sondern auf (nutzerbasierten) Online-Marktplatz-Plattformen wie Amazon veröffentlicht wurden. Diese Plattformen, die in der Regel eine hohe Nutzendenbeteiligung haben, bieten jedoch nur begrenzte Melde- und Beschwerdemöglichkeiten. Damit wird deutlich, dass die Verbreitung von Hass nicht auf wenige Plattformen beschränkt ist, sondern sogar dort gefunden werden kann, wo man sie vielleicht nicht erwarten würde, wie etwa auf Online-Marktplätzen.

Diese Erkenntnisse haben Implikationen für die Governance von Hass im Netz zur Folge: Governance umfasst Handlungen von staatlichen, privaten und zivilgesellschaftlichen Akteur:innen auf verschiedenen Ebenen, einschliesslich der Selbstregulierung durch Plattformen und rechtlicher Massnahmen durch Regierungen (Puppis 2010). Die Erkenntnisse der Studie deuten darauf hin, dass staatliche Massnahmen, die sich ausschliesslich auf grosse Kommunikationsplattformen konzentrieren – wie derzeit in der Schweiz diskutiert – möglicherweise zu kurz greifen. In einer Pressemitteilung des Schweizer Bundesrates vom 5. April 2023 heisst es: „Die Schweizer Bevölkerung soll

mehr Rechte in Bezug auf grosse Kommunikationsplattformen wie Google, Facebook, YouTube und Twitter erhalten.“ Dies umfasst die Möglichkeit, Hassinhalte schnell und einfach bei diesen Plattformen zu melden. Die hohe Relevanz der gespendeten Hassbilder zeigt, dass Nutzende in der Lage sind, unangemessene Kommunikation zu erkennen und zu melden. Daher ist nicht zu erwarten, dass Meldesysteme übermässig missbraucht würden. Die geplanten Massnahmen des Schweizer Bundesrates könnten jedoch Plattformen wie Amazon in der Schweiz nicht betreffen. Auch Plattformen mit geringeren Nutzendenzahlen, die dennoch von erheblicher Bedeutung im Land sind, wie Kleinanzeigenportale, bleiben unberücksichtigt.

Darüber hinaus sollten Hassbotschaften, die sich gegen Personen aufgrund spezifischer Ansichten oder Überzeugungen richten – die nicht durch Artikel 261 des Schweizer Strafgesetzbuchs geschützt sind – zur Überlegung führen, den rechtlichen Rahmen zu erweitern, um diese als potenzielle Ziele von Hassrede zu berücksichtigen.

Die Ergebnisse zeigen auch, dass politische Parteien und Politiker:innen auf die Verwendung von Hassbildern zurückgreifen, um politische Gegner:innen zu diskreditieren, insbesondere im Rahmen negativer Wahlkampagnen. Angesichts des Status und der einzigartigen Rolle dieser Akteur:innen als Vertreter:innen der Demokratie wird angeregt, eine Diskussion darüber zu führen, ob (möglicherweise gesetzliche oder selbst auferlegte) Standards für die Kommunikation von politischen Parteien geschaffen werden sollten, ähnlich den Regeln, die einige Schweizer Parteien für den Einsatz von Künstlicher Intelligenz in Wahlkampagnen verabschiedet haben (Parteien legen Regeln für den Umgang mit Künstlicher Intelligenz fest, 25. September 2023).

Darüber hinaus zeigen die Ergebnisse, dass journalistische Medien eine bedeutende Rolle bei der Verbreitung von Hassbildern spielen, auch wenn sie diese in den begleitenden Artikeln häufig kritisch betrachten. Als Selbstregulierungsmassnahme sollte in den Redaktionen verstärkt über diese Rolle nachgedacht und das Bewusstsein durch die Branchenverbände geschärft werden.

Da visuelle Inhalte in hohem Masse kontextabhängig sind, ist das Entschlüsseln der Botschaft in Bildern oft auf kontextuelles Wissen angewiesen, das je nach Kultur und Zeit variieren kann. Anders als Texten fehlt es visuellen Inhalten in der Regel an leicht identifizierbaren Begriffen oder Akteursnamen, die eine weitere Recherche nach Kontextinformationen ermöglichen könnten. Dies stellt eine besondere Herausforderung

für die Content Moderation dar (Wilson & Land, 2020). Es scheint daher sinnvoll, dass automatisierte Content Moderation durch Personen ergänzt wird, die unterschiedliche Altersgruppen, soziodemografische Merkmale und kulturelle Hintergründe repräsentieren.

Wie jede Studie hat auch dieses Projekt seine Grenzen: Beispielsweise bleibt unklar, inwieweit die gewählte Methode der Datenerhebung als repräsentativ angesehen werden kann. Obwohl dieser Datenspende-Ansatz einige Einblicke in die Verbreitung von Hassbildern auf verschiedenen digitalen Plattformen in der Schweiz liefert, bleibt das Bild dennoch unvollständig. Dies liegt auch daran, dass trotz der grossen Anstrengungen, die in die Citizen-Science-Kampagne gesteckt wurden, nicht genau festgestellt werden kann, wie viele Personen tatsächlich erreicht wurden.

Daher können vier mögliche Erklärungen für die insgesamt geringe Anzahl an eingereichten Hassbildern genannt werden: Erstens war möglicherweise nur wenigen Personen die Möglichkeit zur Teilnahme am Projekt bekannt. Zweitens war die Möglichkeit bekannt, wurde jedoch aufgrund des erforderlichen Aufwands (Hochladen der Bilder und Beantworten von Fragen) nicht genutzt. Drittens war die Möglichkeit zur Teilnahme bekannt, wurde aber aufgrund einer negativen Einstellung gegenüber der Wissenschaft nicht in Anspruch genommen. Viertens war das Forschungsprojekt bekannt, jedoch konnten während der Internetnutzung keine Hassbilder identifiziert werden. Während die ersten drei Erklärungen auf ein Problem im methodischen Ansatz hinweisen würden, würde die vierte Erklärung auf eine relativ gesunde öffentlich vermittelte Kommunikation in der Schweiz hindeuten.

Die Analyse konzentrierte sich ausschliesslich auf Bilder und schloss damit audiovisuelle Inhalte aus, der auf Plattformen wie TikTok und YouTube verbreitet wird. Um diese Lücke zu schliessen, wären ein angepasstes Analysetool und eine entsprechende Datenspende-Plattform für audiovisuelle Inhalte erforderlich. Der Zeitraum für die Datenerhebung war im Februar und März 2023. Angesichts der Tatsache, dass politische Akteur:innen häufig auf Hassbilder zurückgreifen, könnte eine Ausweitung des Zeitrahmens auf den Wahlkampf zusätzliche wertvolle Einblicke liefern.

4 Teil II: Zur Wirkung von Governance-Massnahmen

Der zweite Teil des Berichts fragt nach Möglichkeiten, wie Hassbildern wirkungsvoll begegnet werden kann. Der Bericht folgt dabei einer Governance-Perspektive und trägt damit dem Umstand Rechnung, dass die Eindämmung von (visuellem) Hass im Internet nicht allein staatlicher Regulierung, sondern einer kollaborativen Beteiligung verschiedener involvierter Akteur:innen bedarf. Unter Governance wird mit Mayntz (2004, S. 66) die "Gesamt aller nebeneinander bestehenden Formen der kollektiven Regelung gesellschaftlicher Sachverhalte" verstanden. Die Governance von Hate Speech bezieht sich damit auf die institutionellen, rechtlichen und technischen Massnahmen, die ergriffen werden, um Hassrede zu regulieren, zu kontrollieren und zu reduzieren. Diese Governance umfasst verschiedene Akteure, darunter Regierungen, soziale Medienplattformen, zivilgesellschaftliche Organisationen sowie die Nutzenden selbst (Puppis, 2010). Unterschieden wird zudem zwischen präventiven Massnahmen, die vor der Veröffentlichung von Hassbildern greifen wie bspw. Content Moderation und repressive Massnahmen, die ex post – nach dem Posten des Hassbildes greifen (vgl. Mündges, 2022; Pedrazzi & Oehmer, 2020).

4.1. Forschungsstand, Hypothesen und Subfragestellungen zur Wirkung von Governance-Massnahmen

In der Literatur können folgende Massnahmen zur Governance von Hass identifiziert bzw. abgeleitet werden (vgl. Übersicht in Tabelle 2):

Staatliche Massnahmen: Trotz des internationalen Charakters von Online-Inhalten unterliegen staatliche Massnahmen weiterhin territorialen Grenzen (Schünemann & Steiger, 2023). In vielen europäischen Ländern existieren jedoch vergleichbare gesetzliche Regelungen zur Sanktionierung illegaler Äusserungen. Häufig betreffen diese Gesetze diskriminierende Aussagen, die öffentlich verbreitet werden und sich gegen klar definierte Gruppen, die bestimmte Merkmale wie Geschlecht, Nationalität oder auch Religion teilen, richten. In der Schweiz wurde der Schutzbereich mit der Erweiterung von Artikel 261bis des Strafgesetzbuchs im Jahr 2020 ausgedehnt. Dieser Artikel umfasst nun auch Diskriminierungen aufgrund der sexuellen Orientierung. Darüber hinaus verbietet Artikel 259 des Strafgesetzbuchs das Aufrufen zu Gewalt und die Rechtfertigung von

Verbrechen gegen die Menschlichkeit. Diskriminierungen, die auf persönliche Einstellungen oder den sozialen Status abzielen, bleiben jedoch unberücksichtigt. Jede Bewertung erfordert eine Abwägung zwischen der verfassungsrechtlich garantierten Meinungsfreiheit und dem Schutz vor diskriminierenden oder herabsetzenden Inhalten (Ladeur & Gostomzyk, 2017; Schünemann & Steiger, 2023).

Plattformen verfügen über verschiedene Strategien, um die Verbreitung von Hassbotschaften zu bekämpfen. Eine zentrale Massnahme, die sowohl präventiv als auch repressiv wirken kann, besteht darin, Inhalte zu moderieren. Dabei können Hassbotschaften, die gegen Gesetze oder gegen die eigenen Nutzungsrichtlinien verstossen, entweder manuell oder automatisiert vor ihrer Veröffentlichung blockiert oder nachträglich entfernt werden (Boberg et al., 2018). Neben diesen präventiven Massnahmen stehen Plattformen auch repressive Instrumente zur Verfügung, die gezielt gegen Verfasser:innen von Hassbotschaften eingesetzt werden können. Dazu zählen temporäre Sperrungen oder, als schärfste Massnahme, die dauerhafte Löschung von Accounts. Weitere repressive Ansätze umfassen die algorithmische Depriorisierung oder die Kennzeichnung potenziell problematischer Inhalte, um deren Verbreitung zu begrenzen. Zudem können Plattformen einfache Meldefunktionen bereitstellen, die es Nutzern ermöglichen, Verstösse schnell und unkompliziert zu melden, wodurch die Effizienz der Content Moderation gesteigert werden kann.

Staatliche und zivilgesellschaftliche Akteure können gemeinsam zur Bekämpfung von Hassbotschaften beitragen, indem sie Bildungsangebote entwickeln, die Nutzende darin schulen, Hassinhalte auf digitalen Plattformen zu erkennen und angemessen darauf zu reagieren.

Nutzer:innen können auch durch ihr Verhalten zur Eindämmung solcher Inhalte beitragen (Helberger et al., 2018). Sie haben die Möglichkeit, Plattformfunktionen wie Meldeverfahren zu nutzen, um beleidigende, diskriminierende oder gewaltverherrlichende Inhalte zu melden. Zusätzlich können sie problematische Accounts blockieren oder ihnen entfolgen. Eine weitere Option besteht darin, bei staatlichen Strafverfolgungsbehörden Anzeige zu erstatten, um die Rechtswidrigkeit bestimmter Inhalte prüfen zu lassen. Zudem können sie auf Hassbotschaften direkt reagieren, beispielsweise durch Gegenrede. Diese kann in Form eines Kommentars erfolgen, der die problematischen Aspekte des Inhalts hervorhebt (Hangartner et al.,

2021). Für visuelle Inhalte wäre auch eine visuelle Form der Gegenrede, etwa durch ein Gegenbild, denkbar.

Tabelle 2: Übersicht der Governance-Massnahmen gegen Hass(bilder) im Netz

	Staatliche Akteure	**Plattformen**	**Nutzende**
präventiv	-	- Content Moderation mittels Upload Filter	- Stärkung der Medienkompetenz durch Sensibilisierung auf Folgen von Hate Speech/Hassbildern
repressiv	- Sanktionierung widerrechtlicher Inhalte (bspw. Art. 261 STGB in der Schweiz)	- Content Moderation - Kennzeichnen problematischer Inhalte* (Labeling) - (vorübergehende) Löschung einzelner Posts oder ganzer Accounts - Algorithmische Depriorisierung problematischer Inhalte - Etablieren von Meldemechanismen	- Counter Speech* / Counter-Bild*

Anmerkung: Die mit einem * gekennzeichneten Governance-Massnahmen wurden im Rahmen dieser Studie auf ihre Wirksamkeit überprüft.

In der vorliegenden Studie wird die Wirkung von Governance-Massnahmen geprüft, die ex post – nach dem Posten von Hassbildern – greifen. Die Meinungsäusserungsfreiheit wird durch diese Massnahmen damit weitgehend gewahrt. Zudem werden nur solche Massnahmen geprüft, die den Nutzenden einbeziehen. Folgende Governance-Massnahmen werden geprüft:

a) Kennzeichnung des Inhalts als potenziell verletzend mittels eines Warnhinweises („Labeling“)
b) der Einsatz eines spezifischen Counter-Bild, das direkt auf das Hassbild eingeht (bspw. durch die Verwendung ähnlicher visueller oder textlicher Stilmittel)
c) der Einsatz eines generischen Counter-Bild, das Hass (gegen bestimmte Gruppen) allgemein verurteilt
d) Gegenrede (counter speech) in Textform

4.1.1 Zur Wirkung von Governance-Massnahmen vs. keine Governance-Massnahmen

Studien, die die Wirkung von Governance-Massnahmen von Plattformen und sozialen Medien prüfen, zeigen, dass diese einen Beitrag dazu leisten können, dass falsche oder Hass verbreitende Inhalte keine weitere Reichweite erzielen: So konnte bspw. Mena (2020) feststellen, dass Posts, die einen Hinweis ("labeling") auf einen potenziell falschen Inhalt enthielten, weniger wahrscheinlich verbreitet werden. Hangartner et al. (2021) konnten für die Governance von Hassrede eine Wirkung von empathischer textlicher Gegenrede ausmachen.

Auch wir gehen davon aus, dass Individuen, die den verschiedenen Governance-Massnahmen ausgesetzt sind, wahrscheinlicher dazu beitragen, die Verbreitung des Inhalts einzudämmen als Personen in der Kontrollgruppe. Dieser Argumentation folgend, gehen wir auch davon aus, dass Personen, die Governance-Massnahmen ausgesetzt sind, nicht nur wahrscheinlicher die Verbreitung von Hassbildern minimieren, sondern auch aktiv dagegen Stellung beziehen. Konkret formulieren wir hierzu folgende Hypothese (H):

> **H1:** Individuen, die einer Governance-Massnahme ausgesetzt sind, werden das Hassbild weniger wahrscheinlich **a)** liken, **b)** teilen, **c)** weiterleiten oder **d)** unterstützend kommentieren sowie eher **e)** ablehnend kommentieren oder **f)** melden als diejenigen in der Kontrollgruppe.

4.1.2 Zu den Unterschieden in der Wirksamkeit zwischen Governance-Massnahmen

Wir gehen davon aus, dass nicht alle Governance-Massnahmen eine gleichartige Wirkung entfalten:

> **H2:** Unter Berücksichtigung der Forschung zur visuellen Kommunikation nehmen wir an, dass Individuen, die spezifische oder generische Counter-Hassbilder erhalten, weniger wahrscheinlich Hassbilder **a)** liken, **b)** teilen, **c)** weiterleiten oder **d)** unterstützend kommentieren sowie eher **e)** ablehnend kommentieren oder **f)**

melden als Individuen, die rein textbasierte oder gar keine Governance-Massnahmen erhalten (Carney & Levin, 2002; Leavitt, 2014; Shifman, 2014).

F8: Explorativ ergründen wir, ob spezifische Counter-Hassbilder, die auf ein Hassbild stilistisch oder textlich eingehen wirkungsvoller sind als generische Hassbilder, die allgemein Hass und seine Wirkungen verurteilen.

F9: Zudem ermitteln wir explorativ die Rolle der Quelle einer Intervention, nämlich ob Labeling als Governance-Massnahme einer Plattform wirkungsvoller ist als Counter Speech eines einzelnen Nutzenden.

4.1.3 Zu den Unterschieden in der Wirksamkeit nach Intensität des Hassbildes

Wir nehmen an, dass Governance-Massnahmen in Abhängigkeit der Intensität des verbreiteten Hasses unterschiedlich wirken:

H3: Unter Berücksichtigung des Forschungsstandes zu humorvollen visuellen Inhalten wie beispielsweise Memes, die starke Verbreitung finden (Crawford et al., 2021), gehen wir davon aus, dass Individuen, die humorvollen Hassbildern ausgesetzt sind, unabhängig von Governance-Massnahmen, diese wahrscheinlicher **a)** liken, **b)** teilen, **c)** weiterleiten oder **d)** unterstützend kommentieren beziehungsweise eher **e)** ablehnend kommentieren oder **f)** melden als Individuen, die aggressiven oder zu Gewalt aufrufenden Hassbildern ausgesetzt sind.

4.2 Methode zur Analyse der Wirkungen von Governance-Massnahmen

Um die Wirkung unterschiedlicher Governance-Massnahmen auf die Nutzendeninteraktionen in Zusammenhang mit Hassbildern zum Thema Transgender & Geschlechtsidentität zu untersuchen, haben wir eine standardisierte Befragung im Experimentaldesign durchgeführt. Das in der Medien- und

Kommunikationspolitikforschung selten verwendete experimentelle Design ermöglicht das Testen von Governance-Optionen und Erkenntnisse über deren Folgen (Puppis und Van den Bulck, 2024). Dazu wurde ein 5 x 3 between-subjects faktorielles Design verwendet. Die Teilnehmenden wurden nach dem Zufallsprinzip den Versuchsbedingungen zugewiesen.

Es wurden 5 Governance-Massnahmen getestet:

- keine Governance-Massnahme (Kontrollgruppe)
- Gegenrede in Form eines generischen Gegenbilds als Nutzerkommentar, das ganz allgemein visuell die Unterstützung für Transgender-Personen zum Ausdruck bringt
- Gegenrede in Form eines spezifischen Gegenbilds als Nutzerkommentar, das Elemente des Hassbildes explizit textlich und visuell aufgreift und damit den Inhalt des Hassbildes kontert
- Empathische Gegenrede in Textform als Nutzerkommentar
- Kennzeichnung (Labeling) in Form eines Warnhinweises durch Plattform

Zudem wurden drei Hassbildintensitätsstufen unterschieden (3):

- humorvoll / nicht-aggressiv: Unter diese Form der Diskriminierung fallen negative Darstellung anderer Personen(gruppen), ohne dass dabei auf aggressive (Bild)Sprache (Schimpfwörter, Beleidigungen, entstellende Karikaturen, ...) zurückgegriffen wird. Auch Bilder, die mit dem Stilmittel des Humors, andere Personen(gruppen) ausgrenzen, werden hierunter gefasst.
- aggressiv: Zusätzlich zur negativen Darstellung und Ausgrenzung anderer Personen(gruppen) wird auf eine aggressive (Bild)sprache zurückgegriffen, besipielsweise indem Personen(gruppen) in ihrer Würde herabgesetzt, entmenschlicht, beschimpft oder entwürdigt werden.
- zu Gewalt aufrufend: Diese Ausprägung reicht vom Gutheissen von Gewalt (inklusive Mord und Selbstmord) bis hin zur Mobilisierung und Aktivierung Dritter zur Gewalt.

Die Teilnahme an der Studie war freiwillig. Die Teilnehmenden wurden in der Deutschschweiz durch das Befragungsinstitut Bilendi rekrutiert, das hierfür auf ein bestehendes Panel zurückgegriffen hat. Für die Teilnahme erhielten die Teilnehmenden eine geringe finanzielle Vergütung. Die Datenerhebung fand zwischen dem 21.08.2024 und dem 30.08.2024 statt.

4.2.1 Ethische Erwägungen

Den Teilnehmenden wurde im Verlauf der Befragung je ein Hassbild vorgelegt, das einen negativen Einfluss auf die Gefühlslage der Teilnehmenden haben kann. Als Folge können sie Ärger, Frust oder auch ein Gefühl der Hilfslosigkeit empfinden. Besonders Personen, die auf den Hassbildern diskreditiert werden, können im besonderen Masse davon betroffen sein. Aus diesem Grund befand sich auf der Begrüssungsseite des Fragebogens ein entsprechender Hinweis mit Kontaktinformationen zu Unterstützungsangeboten der Dargebotenen Hand oder zu Beratungsangeboten für Opfer von rassistischer, sexistischer oder queerfeindlicher Diskriminierung der Stadt Bern.

Bevor der Fragebogen ausgefüllt werden konnte, wurden die Teilnehmenden über die Zielstellung der Studie, Ansprechpersonen und den Umgang mit ihren Daten informiert. Zudem wurden sie auch darüber informiert, dass im Fragebogen Hassbilder gezeigt werden, die Einfluss auf ihren Gefühlszustand haben oder sie als Mitglied einer Gruppe beleidigen können. Nur wenn sie damit einverstanden waren und dies entsprechend aktiv bestätigten, konnte mit dem Ausfüllen des Fragebogens begonnen werden. Zudem wurden die Teilnehmenden instruiert, dass sie die Befragung jederzeit abbrechen könnten und wie sie dafür vorzugehen hätten.

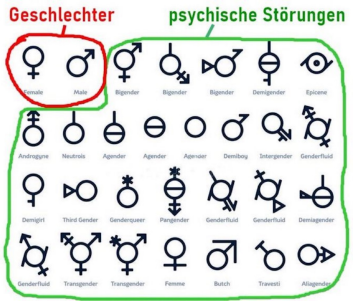
Die Teilnahme wurde auf Personen begrenzt, die mindestens 18 Jahre alt sind. Auf diese Weise sollen Kinder und Jugendliche vor der Rezeption von Hassbildern geschützt werden. Dass die Teilnehmenden dieses Kriterium erfüllen, mussten sie ebenfalls bei der Einverständniserklärung bestätigen. Es wurden keine Kontaktdaten erfasst. Die Teilnehmenden wurden im Verlauf der Befragung gebeten, ihr Alter, ihren Bildungshintergrund und ihre Geschlechtsidentität freiwillig anzugeben. Die gesetzlichen Bestimmungen des Datenschutzes wurden eingehalten. Die Studie wurde durch das Institutional Review Board of the Faculty of Management, Economics and Social

Sciences, University of Fribourg (2024-01-03) genehmigt. Die Studie wurde auf der OSF-Plattform präregistriert (Oehmer-Pedrazzi & Pedrazzi, 14. August 2024).

4.2.2 Ablauf und experimentelle Stimuli

Zu Beginn der Befragung wurden allen Teilnehmenden zunächst zwei in Bezug auf das Untersuchungsobjekt neutrale und über alle Bedingungen hinweg hinsichtlich Inhalts und Nutzendeninteraktionen identische Posts im Design einer Timeline eines sozialen Netzwerks gezeigt. Beide Posts enthielten weder Hassbotschaften noch widmeten sie sich dem Thema Geschlechtsidentität. Die Teilnehmenden mussten für jeden Post angeben, wie wahrscheinlich sie diese Inhalte liken, teilen, extern weiterleiten, unterstützend bzw. ablehnend kommentieren oder melden würden.

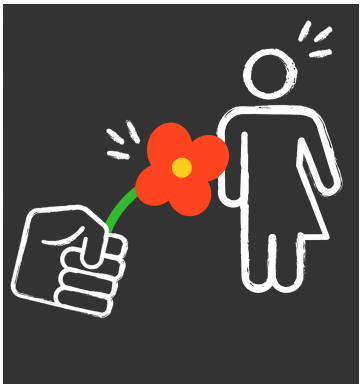
Abbildung 6: im Experiment als Stimuli verwendete Hassbilder unterschiedlicher Intensitäten

humorvoll/nicht-aggressiv	aggressiv	Zu Gewalt aufrufend
		

Es folgte als Stimulus ein weiterer Post, der jeweils eines von drei Hassbildern in unterschiedlicher Intensität (nicht-aggressiv/humorvoll, aggressiv, Gewaltaufruf) enthielt, die sich jeweils gegen eine Transgenderpersonen richteten (vgl. Abbildung 6). Die für das Experiment verwendeten Hassbilder wurden im Rahmen des in Teil I dieses Berichts

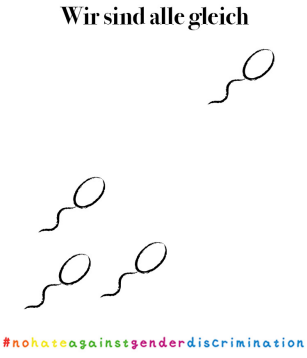
beschriebenen Citizen Science-Ansatzes in der Schweiz ermittelt und deren Hassintensität manuell codiert (siehe ebenso Teil I dieses Berichts).

Abbildung 7: im Experiment verwendete spezifische Counter-Hassbilder

Spezifische Counterbilder		
		

Per Zufall wurden die jeweiligen Hassbilder ohne Governance-Massnahme oder mit einer der vier Governance-Massnahmen – spezifisches Gegenbild (vgl. Abbildung 7), generisches Gegenbild, textlicher Gegenrede oder Warnhinweis (vgl. Abbildung 8) – angezeigt. Die spezifischen und generischen Counter-Hassbilder wurden durch Studierende des Studiengangs Multimedia Production der Fachhochschule Graubünden erstellt.

Abbildung 8: Weitere im Experiment verwendete Governance-Massnahmen

Generisches Counterbild (alle Intensitätsstufen)	Gegenrede in Textform (alle Intensitätsstufen)	Warnhinweis (alle Intensitätsstufen)
	Auch wenn du dir dessen möglicherweise nicht bewusst bist, für inter*, trans* und non-binäre Personen ist dieser Post sehr verletzend.	Der Inhalt dieses Posts könnte andere Personen oder Personengruppen verunglimpfen oder herabsetzen. Erfahre mehr.

Absender des Hassbildes, Datum und Uhrzeit des Posts sowie die Nutzendeninteraktionsmetriken (Likes, Shares, Views) wurden über alle Bedingungen konstant gehalten. In den Bedingungen ohne Governance-Intervention und mit Hinweis durch die Plattform verfügte der Post über keinen Nutzendenkommentar. In den anderen Bedingungen wurden Gegenbilder und Gegenrede von einem für die Studie erfundenen Account namens „no hate @makelovenohate“ als einziger Kommentar auf das Hassbild gepostet, wobei auch hier Datum des Kommentars und Nutzendeninteraktionsmetriken des Kommentars über die Bedingungen konstant gehalten wurden. Damit konnte sichergestellt werden, dass Unterschiede in den Reaktionen auf die unterschiedlichen Governance-Massnahmen zurückgeführt werden können. Die im Experiment dargestellten Verläufe auf der Timeline können Abbildung 9 exemplarisch für ein humorvolles Hassbild entnommen werden.

Wie bei den ersten beiden Posts mussten die Proband:innen die Wahrscheinlichkeit angeben, mit der sie das Hassbild liken, teilen, extern weiterleiten, unterstützend bzw. ablehnend kommentieren oder melden würden. Anschliessend wurden die für die Studie relevanten Kontrollvariablen erhoben bevor der Fragebogen mit einer erneuten Angabe der Kontaktinformationen zu Unterstützungs- und Beratungsangeboten für Opfer von Diskriminierung endete.

4.2.3 Teilnehmende

Um die angemessene Stichprobengrösse zu bestimmen, wurde die Software G*Power (Faul et al., 2009) verwendet. Unser Ziel war es, bei einer angenommenen kleinen bis mittleren Effektgrösse von .175 und bei einer Alpha-Fehlerwahrscheinlichkeit von .05 eine statistische Power von .9 zu erreichen. Dies ergab bei 15 Gruppen eine angestrebte Stichprobengrösse von N=765, was einer Gruppengrösse von n=51 entspricht.

Insgesamt haben 782 Personen den Fragebogen vollständig ausgefüllt. Teilnehmende, die weniger als drei Minuten oder mehr als 15 Minuten für das Ausfüllen des Fragebogens benötigten, wurden ausgeschlossen, so dass N = 666 gültige Fragebögen in die Analyse eingingen. Daraus resultierten variierende Gruppengrössen zwischen n_{Group_min} = 40 und n_{Group_max} = 48. Das Alter der Teilnehmenden reichte von 18 bis 80 Jahren (M = 47.31, SD = 16.42). 51.2 % der Teilnehmenden waren weiblich, 47.9 % männlich, 0.3 % divers und

Abbildung 9: Als Stimuli verwendete Hassbilder und Gegenmassnahmen

Stimulus 1: Hassbild humorvoll, keine Intervention

Stimulus 2: Hassbild humorvoll, generisches Gegenbild als Kommentar

Stimulus 3: Hassbild humorvoll, spezifisches Gegenbild als Kommentar

Stimulus 4: Hassbild humorvoll, Gegenrede als Kommentar

Stimulus 5: Hassbild humorvoll, Warnhinweis durch Plattform

die restlichen 0.6 % verzichteten auf eine Auskunft. 39.9 % der Teilnehmenden hatten zum Zeitpunkt der Teilnahme eine Berufs- oder Hochschulausbildung auf Tertiärniveau absolviert.

4.2.4 Variablen

Abhängige Variablen

Als abhängige Variablen diente die Wahrscheinlichkeit mit der die Teilnehmenden angaben, mit dem Hassbild auf verschiedene Weisen zu interagieren. Dafür wurden die Teilnehmenden gebeten, die Wahrscheinlichkeit von 0 bis 100 Prozent mittels eines Schiebereglers (Ausgangsposition = 0) anzugeben, mit der sie jeweils den Post:

- liken (M = 18.54, SD = 31.07);
- auf dem eigenen Profil teilen (M = 12.23, SD = 24.60);
- Freunden, Verwandten und Bekannten in einer persönlichen Nachricht zukommen lassen (M = 14.73, SD = 26.91);
- in unterstützender Weise kommentieren (M = 9.70, SD = 22.23);
- in ablehnender Weise kommentieren (M = 12.75, SD = 26.21);
- als unangemessen oder unzulässig bei der Plattform melden (M = 16.22, SD = 29.22) würden.

Kontrollvariablen

Wir gehen davon aus, dass die Wirkung der Governance-Massnahmen von verschiedenen individuellen Einstellungen, Kompetenzen, Erfahrungen und Merkmalen der Nutzenden beeinflusst wird. Daher kontrollieren wir zusätzlich zu den Effekten der soziodemographischen Variablen Alter, Geschlecht und Bildung auch für die Effekte der folgenden Kontrollvariablen, die mittels einer 5-Punkte-Likertskala in randomisierter Reihenfolge erhoben wurden:

- Die Einstellung zum Thema Hass im Netz (M = 4.42, SD = .85, Spearman-Brown's ρ = .81) wurde mittels der zwei Items „Hass im Netz ist ein gravierendes Problem“ und „Gegen Hass im Netz muss verstärkt vorgegangen werden“ gemessen, die zu einem Index zusammengefasst wurden. Hohe Werte gehen mit einer grösseren Sensibilität für Hass im Netz als problematisches Phänomen einher.

- Die Einstellung zum Thema Transgender und Geschlechtsidentität (*M* = 3.54, *SD* = 1.10, Cronbach's α = .90) wurde mittels der von den Autor:innen der Studie übersetzten Skala von Walch et al. (2012) erhoben. Die Skala umfasst Items wie „Transgender-Personen sind ein lebendiger und wichtiger Teil unserer Gesellschaft" und „Transgender-Personen sollte es nicht erlaubt sein, mit Kindern zu arbeiten", wobei alle Items so umcodiert wurden, dass hohe Werte einer befürwortenden Einstellung zum Thema Transgender und Geschlechtsidentität entsprechen.
- Die soziale digitale Medienkompetenz (*M* = 4.20, *SD* = .82, Cronbach's α = .84) wurde mittels der Skala von Hoffmann et al. (2019) gemessen. Sie beinhaltet Items wie „Wenn ich im Internet unterwegs bin, achte ich auf meine Ausdrucksweise, wenn ich anderen Nutzenden widerspreche" oder „Wenn ich im Internet unterwegs bin, achte ich darauf, dass ich nicht zu Streitereien oder beleidigenden Diskussionen beitrage". Hohe Werte stehen für einen rücksichtsvollen Umgang sowie eine bewusste Ausdrucks- und Nutzungsweise von digitalen Kommunikationsdiensten.

4.2.5 Datenanalyse

Die Datenanalyse erfolgte mithilfe des Statistikprogramms SPSS. Die Wirkung der einzelnen Governance-Massnahmen wurde mithilfe von Mittelwertvergleichen für unabhängige Stichproben, Kovarianzanalysen (ANCOVA) und Kontrastanalysen sowie für die explorativen Analysen mittels Post-hoc-Tests für Mehrfachvergleiche mit Bonferroni-Korrektur untersucht.

4.3 Ergebnisse zur Wirksamkeit von Governance-Massnahmen

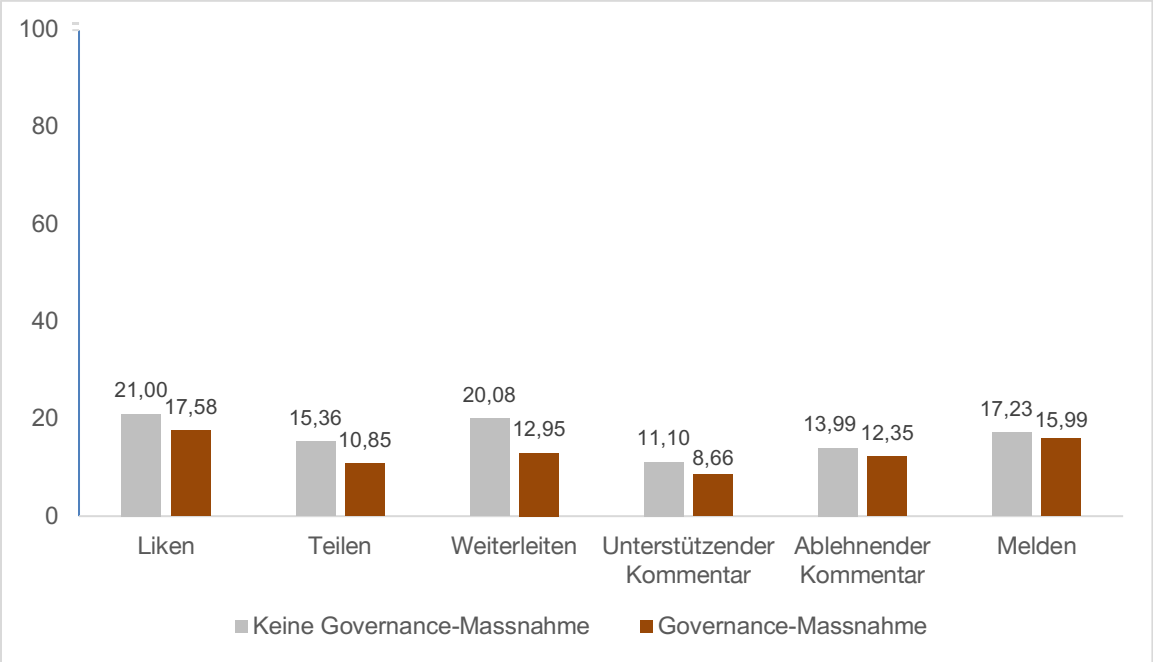
4.3.1 Zur Wirkung von Governance-Massnahmen

Bei Nutzendeninteraktionen, welche die Verbreitung von Hassbildern verstärken oder unterstützen wie bspw. Liken, Teilen, Weiterleiten oder unterstützendes Kommentieren), wurde geprüft, ob durch die Governance-Massnahmen eine reduzierende Wirkung erzielt wird. Umgekehrt wurde bei Nutzendeninteraktionen, die eine Verbreitung eines Hassbildes verhindern oder abschwächen (Melden, ablehnendes Kommentieren), geprüft, ob eine Intervention solches Verhalten fördert. Um die Wirkung der Governance-

Massnahmen auf die Nutzendeninteraktionen zu prüfen, wurden einseitige t-Tests für unabhängige Stichproben durchgeführt.

Die in Abbildung 10 dargestellten Ergebnisse zeigen, dass eine Governance-Massnahme sowohl das Teilen von Hassbildern innerhalb eines sozialen Netzwerks um 4.3 Prozentpunkte ($t(201,43) = 1.75$, $p < .05$, $d = .18$) als auch das Weiterleiten über externe Kommunikationskanäle an Verwandte und Bekannte um 6.8 Prozentpunkte ($t(197,74) = 2.48$, $p < .01$, $d = .25$) verringerte. Bei den anderen untersuchten Governance-Massnahmen zeigte sich keine Wirkung gegenüber der Kontrollbedingung, d.h. der Rezeption des Hassbildes ohne Governance-Massnahme. Somit können die Hypothesen H1b und H1c bestätigt, während H1a, H1d, H1e und H1f verworfen werden.

Abbildung 10: Mittelwerte der Wahrscheinlichkeit einer Nutzendeninteraktion in Bezug auf ein Hassbild (Keine Governace-Massnahme vs. alle Governance-Massnahmen; N=666)



Anmerkung: Die Grafik zeigt die durchschnittliche Wahrscheinlichkeit von 0 bis 100 Prozent, mit der die befragten Personen angaben, mit dem Hassbild zu interagieren.
Lesebeispiel: Personen, die ein der Governance-Massnahmen angezeigt bekommen haben gaben an, das Hassbild mit einer Wahrscheinlichkeit von 17.58 Prozent zu liken. Personen, die nur das Hassbild angezeigt bekommen haben, würde dies mit einer Wahrscheinlichkeit von 21.00 Prozent tun.

4.3.2 Zu den Unterschieden in der Wirksamkeit zwischen Governance-Massnahmen

Nachdem ein Effekt der Governance-Massnahmen auf einzelne Nutzendeninteraktionen festgestellt werden konnte, interessiert nun im Detail, welche konkreten Governance-Massnahmen einen Beitrag dazu leisten können, dass Hassbilder weniger wahrscheinlich geliked, geteilt, an Freunde und Bekannte weitergeleitet oder unterstützend kommentiert werden. Von Interesse waren der Einfluss von textlicher Gegenrede, von Warnhinweisen, von generischen Counter-Hassbildern, die – in diesem Fall – ganz allgemein visuell die Unterstützung für Transgender-Personen zum Ausdruck bringen sowie spezifischen Counter-Hassbildern, die Elemente des Hassbildes explizit textlich und visuell aufgreifen (siehe hierzu Kap. 4.2.2). Eine Governance-Massnahme wäre dann als wirkungsvoll zu betrachten, wenn sie die Wahrscheinlichkeit für eine unterstützende Nutzendeninteraktion reduziert. Die Wirksamkeit der unterschiedlichen Governance-Massnahmen auf die Nutzendeninteraktionen wurde mittels ANCOVAs, jeweils bereinigt um die Effekte der Kontrollvariablen Alter, Geschlecht, Bildung, Einstellung zu Hassrede, Einstellung zu Transgender-Personen und soziale digitale Medienkompetenz, überprüft.

Im Vergleich zu den befragten Personen, die keine Governance-Massnahme erhielten, zeigten die Befragungsteilnehmenden, deren Hassbild mit einer *textlichen Gegenrede* begegnet wurde, die geringsten positiven bzw. verstärkenden Nutzendeninteraktionen: Mit einer 5.82 Prozentpunkte geringeren Wahrscheinlichkeit werden diese Hassbilder geliked, zu 7.94 Prozentpunkten seltener geteilt, beinah zehn Prozentpunkte weniger häufig an Bekannte und Freunde weitergeleitet und 5.37 Prozentpunkte seltener unterstützend kommentiert. Beinah ebenso wirkungsvoll erweisen sich *spezifische Counter-Hassbilder*: Auch sie führen zu einer ebenso geringeren Wahrscheinlichkeit, dass mit ihnen Hassbilder unterstützend oder reichweitenstärkend interagiert wird (vgl. Tabelle 3). Beide Governance-Massnahmen tragen so dazu bei, dass die Reichweite von Hassbildern reduziert werden kann.

Als deutlich weniger wirksam zeigen sich *generische Counter-Hassbilder*: Befragte Personen, die Hassbilder angezeigt bekamen, die mit einem allgemein generisch unterstützenden Counter-Hassbild versehen worden waren, gaben ähnliche Wahrscheinlichkeiten an, mit einem Hassbild positiv unterstützend zu interagieren wie Personen, die nur das Hassbild angezeigt bekommen haben. Besonders überraschend ist, dass generische Counter-Hassbilder sogar die Wahrscheinlichkeit noch erhöhen,

einen positiven Kommentar zu hinterlassen. Dieses Ergebnis ist jedoch nicht signifikant. Auch die *Kennzeichnung des Hassbildes* mit einem Warnhinweis auf mögliche diskriminierende Inhalte führt im Verglich nur zu einer geringeren Interaktionswahrscheinlichkeit als textliche Gegenrede und spezifische Counter-Hassbilder.

Tabelle 3: Wahrscheinlichkeit einer unterstützenden Nutzendeninteraktionen mit einem Hassbild nach Governance-Massnahmen (n=630)

	Liken			Teilen			Weiterleiten			Unterstützender Kommentar		
	M	SE	Δ	M	SE	Δ	M	SE	Δ	M	SE	Δ
Keine Intervention	21.00	2.68		15.36	2.10		20.08	2.31		11.09	1.86	
Gegenbild generisch	20.89	2.74	-0.11	12.15	2.15	-3.21	17.09	2.36	-2.99	11.47	1.90	0.38
Gegenbild spezifisch	15.52	2.70	-5.48	9.41	2.12	**-5.95***	10.83	2.32	**-9.25***	7.15	1.88	-3.94
Gegenrede Text	15.18	2.83	-5.82	7.42[a]	2.22	**-7.94***	10.28	2.43	**-9.80***	5.72	1.97	-5.37
Kennzeichnung	18.63	2.70	-2.37	14.16[a]	2.12	-1.20	13.48	2.33	-6.60	10.12	1.88	-0.97

Anmerkungen: * Kontrastanalyse zeigt einen signifikanten Unterschied ($p < .05$) zur Bedingung ohne Intervention; Bei Werten mit gleichem Superskript [a] wurde bei der Analyse der Kontraste ein signifikanter Unterschied ($p < .05$) gefunden; Die Tabelle zeigt die geschätzten Randmittel M (d.h. um die Effekte der Kontrollvariablen (Einstellung zu Hassrede; Einstellung zu Transgender-Personen; Soziale digitale Medienkompetenz; Alter; Geschlecht; Bildung) bereinigten durchschnittlichen Interaktionswahrscheinlichkeiten; SE = Standardfehler; Δ = Differenz der Randmittel der Governance-Massnahme zur Wahrscheinlichkeit einer Nutzendeninteraktion ohne Governance-Massnahme (=Wirkung der Intervention); grün hinterlegte Felder zeigen Wirkung der Intervention in intendierte Richtung, rot hinterlegte Felder in nicht-intendierte Richtung.
Lesebeispiel: Hassbilder ohne Governance-Massnahme wurden von den befragten Personen mit einer Wahrscheinlichkeit von 20.08 Prozent weitergeleitet. Hassbilder, denen mit einem spezifischen Counter-Hassbild begegnet wurde, wurden mit einer Wahrscheinlichkeit von 10.83 Prozent weitergeleitet.

Zudem wurden auch Effekte auf Nutzendeninteraktionen geprüft, die eine explizite Positionierung gegen das Hassbild deutlich machen: Dazu zählen das öffentliche ablehnende Kommentieren sowie das aktive Melden des Hassbildes. Eine Governance-Massnahme wäre dann als wirksam einzuordnen, wenn sie die Wahrscheinlichkeit für eine ablehnende Nutzendeninterkation erhöht. Auffallend ist zunächst, dass keine der analysierten Governance-Massnahmen zu einer erhöhten Wahrscheinlichkeit führt, dass ein Hassbild ablehnend kommentiert wird (vgl. Tabelle 4). Möglicherweise möchten man hier dem Hassbild – auch nicht mit einem ablehnenden Kommentar – zu einer erhöhten Reichweite verhelfen. Die Wahrscheinlichkeit, ein Hassbild zu melden, wird durch textliche Gegenrede in leichtem Masse erhöht. Auch hier erweist sich Gegenrede in Textform als besonders wirksam. Auffallend ist zudem die niedrigere (allerdings nicht signifikant niedrigere) Wahrscheinlichkeit der Meldung ein Hassbildes bei einem spezifischen Gegenbild. Es lässt sich spekulieren, dass die spezifische Gestaltung der

Intervention dazu führt, dass Nutzende den Eindruck gewinnen, keine weitere Massnahme sei erforderlich – sozusagen, als sei das Problem bereits ausreichend adressiert worden.

Tabelle 4: Wahrscheinlichkeit einer ablehnenden Nutzendeninteraktionen mit einem Hassbild nach Governance-Massnahmen (n=630)

	Ablehnender Kommentar			Melden		
	M	SE	Δ	M	SE	Δ
Keine Intervention	13.98	2.32		17.26	2.52	
Gegenbild generisch	11.26	2.37	-2.72	18.08	2.58	0.82
Gegenbild spezifisch	13.53	2.34	-0.45	10.58	2.54	-6.68
Gegenrede Text	13.16	2.45	-0.82	20.14	2.66	2.88
Kennzeichnung	11.50	2.34	-2.48	15.56	2.54	-1.7

Anmerkungen: Die Tabelle zeigt die geschätzten Randmittel M (d.h. um die Effekte der Kontrollvariablen (Einstellung zu Hassrede; Einstellung zu Transgender-Personen; Soziale digitale Medienkompetenz; Alter; Geschlecht; Bildung) bereinigten durchschnittlichen Interaktionswahrscheinlichkeiten; SE = Standardfehler; Δ = Differenz der Randmittel der Governance-Massnahme zur Wahrscheinlichkeit einer Nutzendeninteraktion ohne Governance-Massnahme (=Wirkung der Intervention) ; grün hinterlegte Felder zeigen Wirkung der Intervention in intendierte Richtung, rot hinterlegte Felder in nicht-intendierte Richtung.
Lesebeispiel: Hassbilder ohne Governance-Massnahme wurden von den befragten Personen mit einer Wahrscheinlichkeit von 17.26 Prozent gemeldet. Hassbilder, denen mit Gegenrede in Textform begegnet wurde, wurden mit einer Wahrscheinlichkeit von 20.14 Prozent gemeldet.

Insgesamt zeigt sich, dass entgegen der Annahme (H2), Gegenrede in visueller Form nicht wirkungsvoller ist als textliche Gegenrede. Im Gegenteil: Empathische Gegenrede zeigt meist im Vergleich die höchste Wirksamkeit – dicht gefolgt von spezifischen Counter-Hassbildern. Die Annahme H2 kann daher nicht beibehalten werden. Das liegt möglicherweise darin begründet, dass Texte zwar weniger stark unsere Aufmerksamkeit binden, sie jedoch weniger Interpretation bedürfen im Vergleich zu visuellen Inhalten. Dies kann wiederum mit einer individuell erhöhten kognitiven Sicherheit einhergehen, was wiederum die erhöhte Wirksamkeit erklären könnte.

Beim direkten Vergleich zwischen generischen und spezifischen Counter-Hassbildern (F8) zeigt sich ein uneinheitliches Bild. Tendenziell erweisen sich spezifische Counterbilder wirksamer bei der Reduktion von unterstützenden und befürwortenden Nutzendeninteraktionen in Bezug auf Hassbilder. Im Gegensatz dazu zeigte sich bei ablehnenden oder die Weiterverbreitung verhindernden Interaktionen eine stärkere Wirkung generischer Gegenbilder. Allerdings erwiesen sich die durch Post-hoc-Tests untersuchten Unterschiede zwischen den beiden Counterbildformaten sowohl für unterstützende als auch verhindernde Nutzendeninteraktionen als nicht signifikant. Eine mögliche Erklärung für die unterschiedlichen Tendenzen liegt darin, dass in der konkreten

Rezeptionssituation spezifische Gegenbilder leichter verständlich und entsprechend als Gegenmassnahme leichter erkennbar sind als generische Gegenbilder. Diese leichtere Informationsverarbeitung führt einerseits zu einer Verringerung der unterstützenden Nutzendeninteraktionen, andererseits kann sie auch zu einer Verringerung ablehnender Nutzendeninteraktionen führen, weil Nutzende den Eindruck gewinnen könnten, dass das Problem dadurch ausreichend adressiert ist.

Zudem machen die Ergebnisse deutlich, dass Warnhinweise, die von Plattformen selbst zur Kennzeichnung problematischer Inhalte eingesetzt werden können, im Vergleich zu nutzendengenerierten textlichen und visuellen Counter-Hassbildern tendenziell weniger wirksam sind (F9). Möglicherweise lässt sich dies mit einer Skepsis gegenüber Content Moderation durch Plattformen und einer erhöhten Glaubwürdigkeit gegenüber anderen Nutzenden erklären (Ozanne et al., 2022).

4.3.3 Zu den Unterschieden in der Wirksamkeit nach Intensität des Hassbildes

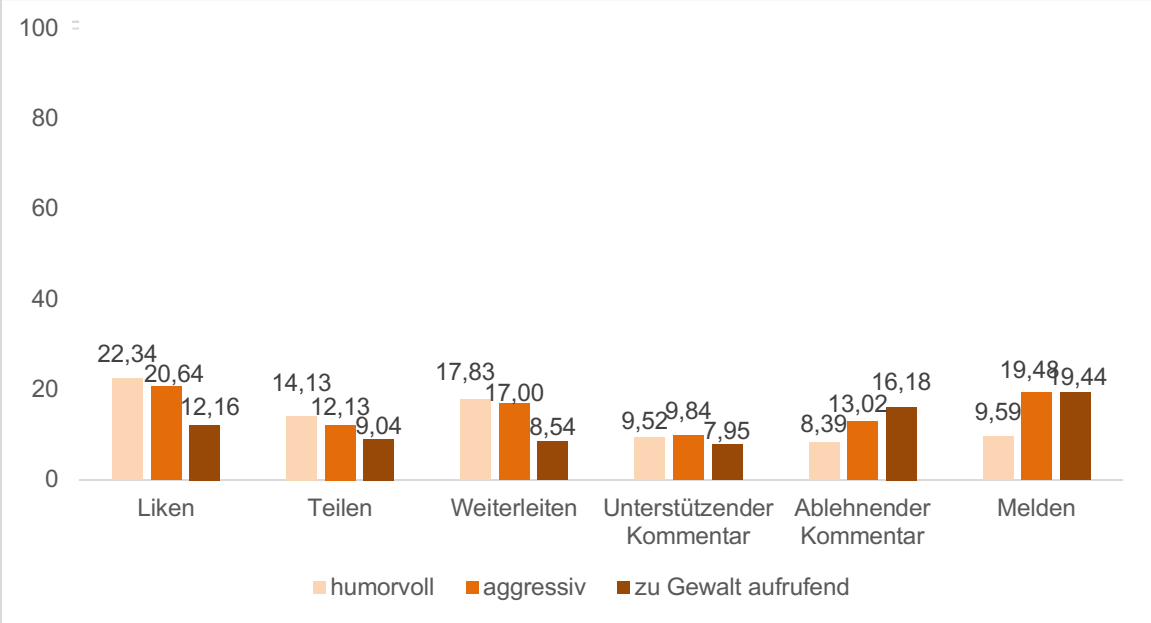
Die analysierten Governance-Massnahmen wirken nicht bei jedem Hassbild gleichermassen: Wir nehmen an, dass die Wirkung von der Intensität eines Hassbildes beeinflusst wird. Humorvolle Hassbilder werden demzufolge wahrscheinlicher geliked, geteilt, an Freunde und Bekannte weitergeleitet und positiv kommentiert als aggressive Hassbilder und Hassbilder, die zu Gewalt aufrufen oder diese darstellen (H3). Letztere werden wahrscheinlicher negativ kommentiert und gemeldet.

Die Wirksamkeit der Governance-Massnahmen auf die Nutzendeninteraktionen Liken, Teilen, Weiterleiten, Kommentieren und Melden wurden wiederum mittels ANCOVAs und Kontraste, jeweils bereinigt um die Effekte der Kontrollvariablen Alter, Geschlecht, Bildung, Einstellung zu Hassrede, Einstellung zu Transgender-Personen und soziale digitale Medienkompetenz überprüft. Die Kontrastanalyse stellt jeweils die statistischen Kennwerte der Intensitätsstufe „humorvoll“, „aggressiv“ und „Gewaltaufruf“ gegenüber.

Erwartungskonform (H3) interagieren die befragten Personen, denen ein humorvolles Hassbild gezeigt wurde, unabhängig von der Governance-Massnahme mit ihnen häufiger in zustimmender Weise: Humorvolle Hassbilder werden signifikant wahrscheinlicher geliked, geteilt und auch an Freunde und Bekannte weitergeleitet als zu Gewalt aufrufende Hassbilder. Beim Liken und externen Weiterleiten zeigte sich auch eine signifikant erhöhte Wahrscheinlichkeit von aggressiven Hassbildern gegenüber zu Gewalt

aufrufenden Hassbildern. Zudem werden humorvolle Hassbilder mit einer geringeren Wahrscheinlichkeit gemeldet als aggressive oder zu Gewalt aufrufende Hassbilder. Das schädigende Potenzial wird hier möglicherweise als geringer angesehen und daher eine Meldung für weniger notwendig erachtet. Mit zunehmender Intensität eines Hassbildes, dem die Teilnehmenden des Experiments ausgesetzt wurden, reagierten diese mit abnehmender Wahrscheinlichkeit mit einer zustimmenden Nutzendeninteraktion. Zudem wurden aggressive und zu Gewalt aufrufende Hassbilder tendenziell auch wahrscheinlicher mit einem ablehnenden und weniger wahrscheinlich mit einem unterstützenden Kommentar versehen (vgl. Abbildung 11).

Abbildung 11: Mittelwerte der Wahrscheinlichkeit einer Nutzendeninteraktion mit einem Hassbild nach Intensität des Hassbildes (n=630)



Anmerkung: Die Grafik zeigt die geschätzten Randmittel (d.h. um die Effekte der Kontrollvariablen (Einstellung zu Hassrede; Einstellung zu Transgender-Personen; Soziale digitale Medienkompetenz; Alter; Geschlecht; Bildung) bereinigten durchschnittlichen Interaktionswahrscheinlichkeiten von 0 bis 100 Prozent mit der die befragten Personen angaben, mit dem Hassbild zu interagieren – systematisiert nach der Intensität des Hassbildes.
Lesebeispiel: Personen, die ein humorvolles Hassbild erhielten, würden dies mit einer Wahrscheinlichkeit von 22.34 Prozent liken; Personen, die ein zu Gewalt aufrufenden Hassbild erhielten, würden dies nur mit einer Wahrscheinlichkeit von 12.16 Prozent tun.

Zusätzlich wurden explorativ auch Unterschiede in der Wirksamkeit der einzelnen zustimmenden, verstärkenden (vgl. Tabelle 5) sowie ablehnenden und abschwächenden

(vgl. Tabelle 6) Governance-Massnahmen innerhalb der sowie nach Intensitätsstufen mittels Post-hoc-Tests mit Bonferroni-Korrektur für Mehrfachvergleiche untersucht.

Die Ergebnisse deuten darauf hin, dass spezifische Counter-Hassbilder, die Merkmale des Originalhassbildes aufgreifen, bei humorvollen sowie bei aggressiven Hassbildern vergleichsweise stärker wirken als andere Governance-Massnahmen: Sie führen dazu, dass ein humorvolles oder aggressives Hassbild weniger wahrscheinlich geliked, geteilt, an Freunde und Bekannte weitergeleitet oder unterstützend kommentiert wird.

Gegenrede in Textform erweist sich über alle Intensitätsstufen hinweg und in Bezug auf alle Nutzendeninteraktionen als reduzierend, insbesondere auch bei zu Gewalt aufrufenden Hassbildern, wo spezifische Counterbilder eine geringere Wirkung zeigen. Warnhinweise erweisen sich insbesondere bei Gewalt verherrlichenden und damit besonders intensiven Hassbildern als wirksam. Sie tragen jeweils zur Reduktion wahrscheinlicher Nutzendeninteraktionen bei und verfügen damit über das Potenzial, die Reichweite solcher Hassbilder zu verringern.

Bei den Ergebnissen zur Kennzeichnung stechen die inkonsistenten und sogar gegenüber der Bedingung ohne Intervention erhöhten Interaktionswahrscheinlichkeiten bei aggressiven Hassbildern hervor. Hier kann spekuliert werden, dass die Kennzeichnung des Hassbildes mit einem Warnhinweis zu einer Überkorrektur durch die Teilnehmenden geführt hat, bedarf aber zur Klärung weiterführender – auch auf qualitativen Ansätzen basierender – Forschung.

Insgesamt am wenigsten wirksam erscheinen generische Counterbilder, insbesondere bei Hassbildern höherer Intensitätsstufen, wo sie auf Nutzendeninteraktionen eher verstärkend wirken. Hier lässt sich spekulieren, dass die generell zu mehr Toleranz aufrufende Botschaft der generischen Gegenbilder bei gewissen Personen Reaktanz auslösen kann, dies bedarf weiterer Forschung. Widersprüchlich erscheint auch der Befund, dass sowohl generische als auch spezifische Counter-Hassbilder die Wahrscheinlichkeit, ein zu Gewalt aufrufendes Hassbild zu teilen, verringern, sie aber wahrscheinlicher dazu führen, dass ein solches geliked wird. Auch hier bedarf es weiterführender - auch qualitativer – Ansätze, um dieses Muster erklären zu können.

Tabelle 5: Wahrscheinlichkeit einer zustimmenden Nutzendeninteraktionen mit einem Hassbild nach Governance-Massnahmen und Hassbildintensitätsstufen (n=630)

	Liken			Teilen			Weiterleiten			Unterstützender Kommentar		
	M	SE	Δ	M	SE	Δ	M	SE	Δ	M	SE	Δ
Humorvolles / nicht-aggressives Hassbild												
Keine Intervention	30.86	4.67		22.16	3.69		30.87	4.02		11.79	3.30	
Gegenbild generisch	22.51	4.92	-8.35	11.62	3.89	-10.54	16.56	4.23	-14.31	12.14	3.48	0.35
Gegenbild spezifisch	15.45	4.80	-15.41	8.97	3.80	-13.19	12.56	4.13	**-18.31***	8.11	3.40	-3.68
Gegenrede Text	18.30	4.97	-12.56	9.72	3.93	-12.44	13.82	4.28	**-17.05***	5.02	3.52	-6.77
Kennzeichnung	24.60	4.85	-6.26	18.18	3.84	-3.98	15.32	4.17	-15.55	10.56	3.43	-1.23
Aggressives Hassbild												
Keine Intervention	21.00	4.59		10.55	3.63		19.10	3.95		11.35	3.25	
Gegenbild generisch	22.92	4.80	1.92	13.60	3.79	3.05	20.72	4.13	1.62	9.49	3.39	-1.86
Gegenbild spezifisch	14.23	4.46	-6.77	8.65	3.53	-1.90	11.24	3.83	-7.86	6.08	3.15	-5.27
Gegenrede Text	17.59	4.78	-3.41	9.24	3.78	-1.31	13.19	4.11	-5.91	8.61	3.38	-2.74
Kennzeichnung	27.47	4.52	6.47	18.64	3.58	8.09	20.74	3.89	1.64	13.67	3.20	2.32
Zu Gewalt aufrufendes Hassbild												
Keine Intervention	11.91	4.50		13.71	3.56		11.11	3.87		13.67	3.20	
Gegenbild generisch	17.79	4.41	5.88	11.31	3.49	-2.40	14.36	3.79	3.25	10.21	3.18	-3.46
Gegenbild spezifisch	16.83	4.61	4.92	10.64	3.65	-3.07	8.67	3.97	-2.44	12.58	3.12	-1.09
Gegenrede Text	9.97	4.80	-1.94	3.49	3.79	-10.22	4.11	4.13	-7.00	7.38	3.26	-6.29
Kennzeichnung	4.33	4.56	-7.58	6.07	3.60	-7.64	4.43	3.92	-6.68	3.52	3.39	-10.15

Anmerkungen: * Kontrastanalyse zeigt einen signifikanten Unterschied ($p < .05$) zur Bedingung ohne Intervention; Die Tabelle zeigt die geschätzten Randmittel M (d.h. um die Effekte der Kontrollvariablen (Einstellung zu Hassrede; Einstellung zu Transgender-Personen; Soziale digitale Medienkompetenz; Alter; Geschlecht; Bildung) bereinigten durchschnittlichen Interaktionswahrscheinlichkeiten; SE = Standardfehler; Δ = Differenz der Randmittel der Governance-Massnahme zur Wahrscheinlichkeit einer Nutzendeninteraktion ohne Governance-Massnahme (=Wirkung der Intervention) ; grün hinterlegte Felder zeigen Wirkung der Intervention in intendierte Richtung, rot hinterlegte Felder in nicht-intendierte Richtung.
Lesebeispiel: Humorvolle Hassbilder ohne Governance-Massnahme wurden von den befragten Personen mit einer Wahrscheinlichkeit von 30.86 Prozent geliked. Humorvolle Hassbilder, denen mit einem spezifischen Counter-Hassbild begegnet wurde, wurden mit einer Wahrscheinlichkeit von 15.45 Prozent geliked.

Im Gegensatz zu den Ergebnissen in Zusammenhang mit verstärkenden Nutzendeninteraktionen lässt sich für abschwächende Nutzendeninteraktionen über die Intensitätsstufen kein konsistentes Muster für die unterschiedlichen Governance-Massnahmen identifizieren (vgl. Tabelle 6). Insgesamt deuten die Werte darauf hin, dass Nutzende auch ohne Intervention in der Lage sind, Hassbilder als solche zu erkennen und der Intensität eines Hassbildes entsprechend zu reagieren. Es lässt sich spekulieren, dass die Governance-Massnahmen sowohl bezüglich ablehnenden Kommentierens als auch Meldens den Eindruck erwecken können, dass keine weiteren Aktionen nötig sind, was eine Erklärung für die zahlreichen negativen Unterschiede im Vergleich zur Bedingung ohne Intervention liefern könnte.

Tabelle 6: Wahrscheinlichkeit einer ablehnenden Nutzendeninteraktionen mit einem Hassbild nach Governance-Massnahmen und Hassbildintensitätsstufen (n=630)

	Ablehnender Kommentar			Melden		
	M	SE	Δ	M	SE	Δ
Humorvolles / nicht-aggressives Hassbild						
Keine Intervention	10.42	4.07		8.79	4.38	
Gegenbild generisch	3.85	4.29	-6.57	9.51	4.61	0.72
Gegenbild spezifisch	13.43	4.19	3.01	6.91	4.50	-1.88
Gegenrede Text	4.20	4.34	-6.22	14.38	4.67	5.59
Kennzeichnung	10.06	4.23	-0.36	8.34	4.55	-0.45
Aggressives Hassbild						
Keine Intervention	12.26	4.01		18.73	4.31	
Gegenbild generisch	12.39	4.19	0.13	28.79[c]	4.50	10.06
Gegenbild spezifisch	14.56	3.89	2.30	15.24	4.18	-3.49
Gegenrede Text	18.32	4.17	6.06	23.49	4.49	4.76
Kennzeichnung	7.53	3.95	-4.73	11.16[c]	4.24	-7.57
Zu Gewalt aufrufendes Hassbild						
Keine Intervention	18.92	3.93		23.69	4.22	
Gegenbild generisch	16.31	3.85	-2.61	15.94	4.13	-7.75
Gegenbild spezifisch	12.67	4.02	-6.25	9.20[d]	4.33	-14.49
Gegenrede Text	16.27	4.19	-2.65	22.00	4.5	-1.69
Kennzeichnung	16.74	3.98	-2.18	26.37[d]	4.27	2.68

Die Tabelle zeigt die geschätzten Randmittel M (d.h. um die Effekte der Kontrollvariablen (Einstellung zu Hassrede; Einstellung zu Transgender-Personen; Soziale digitale Medienkompetenz; Alter; Geschlecht; Bildung) bereinigten durchschnittlichen Interaktionswahrscheinlichkeiten; SE = Standardfehler; Δ = Differenz der Randmittel der Governance-Massnahme zur Wahrscheinlichkeit einer Nutzendeninteraktion ohne Governance-Massnahme (=Wirkung der Intervention) ; grün hinterlegte Felder zeigen Wirkung der Intervention in intendierte Richtung, rot hinterlegte Felder in nicht-intendierte Richtung.
Lesebeispiel: Humorvolle Hassbilder ohne Governance-Massnahme wurden von den befragten Personen mit einer Wahrscheinlichkeit von 8.79 Prozent gemeldet. Humorvolle Hassbilder, denen mit Gegenrede in Textform begegnet wurde, wurden mit einer Wahrscheinlichkeit von 14.38 Prozent gemeldet.

Bei den ablehnenden Kommentaren zeigen die spezifischen Gegenbilder bei humorvollen und aggressiven Hassbildern, aber nicht bei zu Gewalt aufrufenden Hassbildern, eine Wirkung. Hier wäre mit weiterer Forschung zu prüfen, ob diese Unterschiede eher auf inhaltliche und gestalterische Spezifika der Gegenbilder zurückzuführen sind als auf die Intensitätsstufen.

Was das Melden angeht, so erweist sich empathische Gegenrede am ehesten wirksam. Bei humorvollen und aggressiven Hassbildern führt sie, womöglich dank des geringen Interpretationsspielraums, zu einer erhöhten Aufmerksamkeit, dass solche Inhalte

verletzend sein können, was wiederum zur Folge haben kann, dass solche Hassbilder wahrscheinlicher gemeldet werden.

4.4 Fazit & Implikationen für die Governance

Der zweite Teil des vorliegenden Berichts befasste sich mit der Wirkung einzelner Governance-Massnahmen von Hassbildern. Im Fokus standen dabei v.a. Massnahmen, die auf eine Umsetzung durch oder Berücksichtigung von Nutzenden angewiesen sind. Zudem wurden nur Governance-Massnahmen getestet, die ex post und damit nach der Veröffentlichung eines Hassbildes greifen und keinen etwaigen Eingriff in die Meinungsäusserungsfreiheit darstellen. Geprüft wurden die Effekte der Governance-Massnahmen Gegenrede, generische Counter-Hassbilder, spezifische Counter-Hassbilder sowie Warnhinweise auf die Nutzendeninteraktionen. Eine Governance-Massnahme gilt dann als wirksam, wenn sie einerseits geeignet ist, unterstützende und reichweitenverstärkende Interaktionen (Liken, Teilen, Weiterleiten und positiv kommentieren) mit einem Hassbild einzudämmen. Andererseits gilt sie auch als wirksam, wenn sie ablehnende Nutzendeninteraktionen (melden, negativ kommentieren) mit dem Hassbild fördert. Die Wirksamkeit wurde experimentell geprüft.

Ganz allgemein lässt sich feststellen, dass Governance-Massnahmen bei Hassbildern eine Wirkung zeigen: Die Befragten, denen eine der vier Governance-Massnahmen gezeigt wurde, haben wahrscheinlicher vom Teilen oder privaten Weiterleiten des Hassbildes abgesehen als Personen, die keine Governance-Massnahme angezeigt bekommen haben. Eine weitere durch Nutzende oder algorithmisch getriebene Verbreitung und damit auch eine Erhöhung der Reichweite der Hassbilder innerhalb wie auch ausserhalb eines sozialen Netzwerks, in dem ein Hassbild zirkuliert, kann damit eingedämmt werden. Dieses Ergebnis ist von besonderer Bedeutung, wenn man bedenkt, dass ein extern an Freunde und Bekannte weitergeleitetes Hassbild von eine:r persönlich nahe:n Absender:in stammt, der:ie in digitalen Kommunikationsprozessen über hohes Vertrauen verfügen (Metzger et al., 2010).

Als besonders wirkungsvoll erwiesen sich eine empathische Form textlicher Gegenrede sowie spezifische Counter-Hassbilder, die visuelle Merkmale des Hassbildes explizit wieder aufgegriffen haben. Beide Massnahmen sind geeignet, die Reichweite von Hassbildern zu reduzieren, indem sie die Wahrscheinlichkeit, dass Hassbilder geliked,

geteilt, oder an Freund:innen und Bekannte weitergeleitet werden, senken. Zudem verringern sie auch, dass Hassbilder positiv kommentiert werden.

Entgegen der Annahme sind visuelle Counter-Hassbilder nicht wirkungsvoller als textliche Gegenrede. Hassbildern kann daher auch mit empathischer Gegenrede (die Ergebnisse von Hangartner et al. (2021) für Hassrede bestätigend) wirkungsvoll begegnet werden. Warnhinweise vor Hassbildern mit einem Verweis auf einen möglichen diskriminierenden Inhalt zeigen im Vergleich zu Gegenrede in textlicher oder visueller Form durch Nutzendenkommentaren eine geringere Wirkung.

Berücksichtigt man die Intensitätsstufen von Hassbildern (von humorvoll über aggressiv zu Gewaltaufruf), so zeigen sich unterschiedliche Wirkmechanismen: Während Gegenrede in Textform über alle Intensitätsstufen hinweg und bei humorvollen und aggressiven Hassbildern vor allem spezifische Counter-Hassbilder zur Verringerung der Reichweite beitragen können, erweisen sich bei Gewalt verherrlichenden Hassbildern Warnhinweise auf den potenziell diskriminierenden Inhalt durch die Plattformbetreiber als vergleichsweise effektiv. Widersprüchlich sind hingegen die Ergebnisse zum Einfluss von Warnhinweisen bei aggressiven Hassbildern: Hier kann eine Kennzeichnung durch Plattformen sogar zu einer Verstärkung der Reichweite führen. Am wenigsten effektiv in Bezug auf die Verringerung von Nutzendeninteraktionen, die zu einer verstärkten Verbreitung von Hassbildern führen können, erwiesen sich generische Counterbilder.

Für die Governance von Hassbildern lassen sich daraus folgende Implikationen ableiten: Es zeigt sich, dass Nutzende bis zu einem bestimmen Mass selbst die Möglichkeit haben, einen Einfluss auf die Reichweite von Hassbildern auszuüben: Auf textlicher empathischer Gegenrede und spezifischen Counter-Hassbildern basierende Nutzendenkommentare als Reaktion auf Hassbilder senken die Wahrscheinlichkeit, mit der ein Hassbild geliked, geteilt oder weitergeleitet wird. Die Reichweite von diskriminierenden Bildern kann somit als Folge von nutzendengenerierten Interaktionen gesenkt werden, was wiederum die algorithmische Verbreitung von Inhalten beeinflusst, die in der Regel Inhalte mit hoher Interaktionsrate zusätzlich verstärkt. In der Folge kann so ein Beitrag zu einem gesünderen und weniger diskriminierenden Online-Diskurs geleistet werden.

Um Nutzende darin zu ermächtigen, verstärkt auf Hassbilder mit Gegenrede oder spezifischen Counter-Hassbildern zu reagieren, müssen zwei Voraussetzungen erfüllt

sein: Erstens müssten die Plattformen im Rahmen ihres Designs und ihrer Architektur Nutzenden die Möglichkeit einräumen, auf Inhalte mit Kommentaren in Form von Texten und Bildern zu reagieren. Auf einigen Plattformen (wie bspw. der stark visuell orientierten Plattform Instagram) kann nicht mit visuellen Inhalten kommentiert werden. Hassbildern kann somit nicht mit einem Counter-Hassbild unmittelbar entgegnet werden. Zweitens, müssen Nutzende über die Wirkmacht ihrer Nutzendeninteraktionen im Rahmen von (Weiter)Bildungsmassnahmen hingewiesen werden.

Anzudenken wäre bspw. auch die Initiierung oder Förderung von unabhängigen Organisationen, die sich aktiv als «trusted commentators» mit Gegenrede sowie spezifischen Counter-Hassbildern in den Onlinediskurs einbringen. Eine solche Leistung könnte auch als Service Public qualifiziert werden und die hierfür notwendigen finanziellen Ressourcen entsprechend aus einer Kombination von öffentlichen und privaten Förderinstitutionen sowie den Plattformen selbst erbracht werden. Kontroll- und Beteiligungsmechanismen müssten dabei jedoch sicherstellen, dass die Organisationen im Sinne der Allgemeinheit agieren und nicht partikularen Interessen folgen.

Warnhinweise, die von Plattformen im Rahmen der Content Moderation einzelnen Posts angefügt werden, zeigen vor allem Wirkung bei Hassbildern, die zu Gewalt aufrufen oder Gewalt verherrlichen. Bei Hassbildern von geringerer Intensität führen sie teilweise sogar zu einem gegenteiligen Effekt. Möglicherweise wird dies als zu starke Einmischung oder Bevormundung durch die Plattform wahrgenommen oder lässt sich auf ein geringes Vertrauen in automatisiert betriebener Content-Moderation zurückführen (Ozanne et al., 2022). Auf der Basis dieser Erkenntnisse wäre die Empfehlung für einen zurückhaltenden Einsatz von Warnhinweisen durch Plattformen und eine deutliche Beschränkung bei gravierenden Formen visuellen Hasses auszusprechen. In diesem Zusammenhang ist auch auf die Komplexität automatisierter Content Moderation, trotz aktueller Fortschritte in KI-betriebener Computer Vision, beim (interkulturellen) Verständnis von Bildern und deren Zusammenspiel mit Text zu verweisen (Cheema et al., 2023).

An dieser Stelle ist in Bezug auf das Untersuchungsdesign limitierend anzuführen, dass die Teilnehmenden im Rahmen des Experiments lediglich nach ihrer Absicht zu Interaktionen wie Likes, Teilen, usw. einzelner, isolierter Posts befragt wurden. Dieses Vorgehen spiegelt jedoch nicht zwangsläufig ihr tatsächliches Verhalten in einer konkreten Nutzungssituation wider. Dennoch bietet ein solches Verfahren eine

wahrscheinliche Annäherung an das tatsächliche Verhalten (Mosleh et al., 2020). Es sei des weiteren angemerkt, dass die hier untersuchten Wirkmechanismen auf Governance-Massnahmen basieren, die an drei ausgewählten Hassbildern gegen Transgender-Personen getestet wurden. Für eine höhere Belastbarkeit der gefundenen Befunde wäre es daher wünschenswert zu prüfen, ob die gefundenen Wirkmechanismen auch auf Hassbilder zutreffen, die andere Gruppen diskriminieren. Zudem wäre es erstrebenswert auch den Korpus getesteter Hassbilder und Hassbilderintensitäten zu vergrössern, um ausschliessen zu können, dass die gefundenen Zusammenhänge rein bildspezifisch sind. Hierzu würde sich zusätzlich zu weiteren Experimenten eine Erweiterung des Studiendesigns um einen qualitativen Ansatz anbieten. So könnten bspw. im Rahmen von Interviews oder Fokusgruppen Erkenntnisse zum Umgang mit Hassbildern sowie zur Wahrnehmung der Governance-Massnahmen gewonnen werden. Zudem wäre es auch möglich, die Befragung im Experimentaldesign um einen Ansatz des «Lauten Denkens» (Bilandzic, 2017) zu ergänzen, mit dem die befragten Personen nicht nur den Fragebogen ausfüllen, sondern auch laut kommunizieren, was Sie beim Anblick der Posts und der Auswahlmöglichkeiten zum Umgang mit diesen Posts denken.

Literatur

Askanius, T. (2021). On Frogs, Monkeys, and Execution Memes: Exploring the Humor-Hate Nexus at the Intersection of Neo-Nazi and Alt-Right Movements in Sweden. Television & New Media, 22(2), 147-165. https://doi.org/10.1177/1527476420982234

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, 27.

Bilandzic, H. (2017). Lautes Denken. In L. Mikos & C. Wegener (Hrsg.), *Qualitative Medienforschung. Ein Handbuch* (S. 406–413). UVK.

Bilewicz, M. & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. Political Psychology, 41(1), 3–33. https://doi.org/10.1111/pops.12670

Blaya, C., Audrin, C. & Skrzypiec G. (2020). School bullying, perpetration and cyberhate: overlapping issues. Contemporary School Psychology. Advance online publication. https://doi.org/10.1007/s40688-020-00318-5

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

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. Law and Philosophy, 36(5), 419–468. https://doi.org/10.1007/s10982-017-9297-1

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. https://doi.org/10.1023/A:1013176309260

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. https://doi.org/10.1016/j.avb.2021.101608

Cheema, G. S., Hakimov, S., Müller-Budack, E., Otto, C., Bateman, J. A., & Ewerth, R. (2023). Understanding image-text relations and news values for multimodal news analysis [Hypothesis and Theory]. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1125533

Crawford, B., Keen, F. & Suarez-Tangil, G. (2021). Memes, Radicalisation, and the Promotion of Violence on Chan Sites. Proceedings of the International AAAI Conference on Web and Social Media, 15, 982–991. https://doi.org/10.1609/icwsm.v15i1.18121

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Frischlich, L. (2018). „Propaganda 3 "–Einblicke in die Inszenierung und Wirkung von Online-Propaganda auf der Makro-Meso-Mikro-Ebene [Propaganda 3 – Insights into the staging

and impact of online propaganda at the macro-meso-micro level]. In Fake News, Hashtags & Social Bots: Neue Methoden populistischer Propaganda (pp. 133-170).

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

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

Graber, R. & Lindemann, T. (2018). Neue Propaganda im Internet. Social Bots und das Prinzip sozialer Bewährtheit als Instrumente der Propaganda [New Propaganda on the Internet. Social Bots and the Principle of Social Proof as Instruments of Propaganda]. In Fake News, Hashtags & Social Bots: New Methods of Populist Propaganda (pp. 51-68).

Grosse Kommunikationsplattformen: Bundesrat strebt Regulierung an ["Major Communication Platforms: Federal Council Aims for Regulation] (05. April 2023). https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-94116.html

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

Hoffmann, C. P., Weber, J., Zepic, R., Greger, V., & Krcmar, H. (2019). Dimensionen digitaler Mündigkeit und politische Beteiligung im Netz. In I. Engelmann, M. Legrand, & H. Marzinkowski (Eds.), *Politische Partizipation im Medienwandel* (pp. 79-99). https://doi.org/10.17174/dcr.v6.4Leavitt, A (2014) From #FollowFriday to YOLO: Exploring the cultural salience of Twitter Memes. In: Weller, K, Bruns, A, Burgess, J, Mahrt, M, Puschmann, C (eds) Twitter and Society. New York: Peter Lang, pp. 137–154.

Hornuff, D. (2020). Hassbilder. Gewalt posten, Erniedrigung liken, Feindschaft teilen. [Hassbilder. Posting Violence, Liking Humiliation, Sharing Hostility.] Berlin: Wagenbach.

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

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 Soc Change* 69, 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

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.

Künzler, M. (2013). *Mediensystem Schweiz.* Konstanz, München: UVK Verlagsgesellschaft.

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.

Leavitt, A (2014) From #FollowFriday to YOLO: Exploring the cultural salience of Twitter Memes. In: Weller, K, Bruns, A, Burgess, J, Mahrt, M, Puschmann, C (eds) Twitter and Society. New York: Peter Lang, pp. 137–154.

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* (pp. 65-76). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90171-8_4

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. Policy & internet, 12(2), 165-183.

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

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, *60*(3), 413-439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS ONE 15*(2): e0228882. https://doi.org/10.1371/journal.pone.0228882

Mündges, S. (2022). Die Verantwortung von Facebook & Co. Über den Umgang digitaler Plattformen mit Hate Speech. In: Weitzel, G., Mündges, S. (eds) Hate Speech. Aktivismus- und Propagandaforschung (231- 248). Springer VS. https://doi.org/10.1007/978- 3- 658- 35658- 3_12

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

Näsi, M., Räsänen, P., Hawdon, J., Holkeri E. & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. Information Technology & People, 28(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

Oehmer-Pedrazzi, F., & Pedrazzi, S. (2024, August 14). Governance of Visual Hate Images on the Internet. https://doi.org/10.17605/OSF.IO/URBVS

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. Big Data & Society, 9(2). https://doi.org/10.1177/20539517221115666

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.

Paciello, M., D'Errico, F., Saleri, G. & Lamponi, E. (2021). Online sexist meme and its effects on moral and emotional processes in social media. Computers in Human Behavior, 116, 106655.

Parteien geben sich Regeln im Umgang mit Künstlicher Intelligenz [Political Parties Establish Rules for Dealing with Artificial Intelligence] (25. September 2023). https://www.swissinfo.ch/ger/alle-news-in-kuerze/parteien-geben-sich-regeln-im-umgang-mit-kuenstlicher-intelligenz/48836884

Pedrazzi, S., & Oehmer, F. (2020). Communication rights for social bots? Options for the governance of automated computer-generated online identities. Journal of Information Policy, 10, 549-581. https://doi.org/10.5325/jinfopoli.10.2020.0549

Puppis, M. (2010). Media Governance: A New Concept for the Analysis of Media Policy and Regulation. Communication, Culture & Critique, 3(2), 134-149. https://doi.org/10.1111/j.1753-9137.2010.01063.x

Puppis, M., & Van den Bulck, H. (2024). Methods for Global Media and Communication Governance Research. In C. Padovani, V. Wavre, A. Hintz, G. Goggin, & P. Iosifidis (Eds.), *Global Communication Governance at the Crossroads* (pp. 371-387). Springer International Publishing. https://doi.org/10.1007/978-3-031-29616-1_21

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M. & Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young Finnish Facebook users. Violence and Victims, 31(4), 708–726. https://doi.org/10.1891/0886-6708.VV-D-14-00079

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). https://doi.org/10.1177/20563051211052906

Sakki, I. & Castrén, L. (2022). Dehumanization through humour and conspiracies in online hate towards Chinese people during the COVID-19 pandemic. *British Journal of Social Psychology*, *61*(4), 1418-1438. https://doi.org/10.1111/bjso.12543

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, 0(0), 1–19. 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, 0(0). https://doi.org/10.1177/14614448231198169

Schmitt, J. B., Harles, D. & Rieger, D. (2020). Themen, Motive und Mainstreaming in rechtsextremen Online-Memes. *M&K Medien & Kommunikationswissenschaft, 68*(1–2), 73–93. https://doi.org/10.5771/1615-634X-2020-1-2-73

Schünemann, W.J. & Steiger, S. (2023). Die Regulierung von Internetinhalten am Beispiel Hassrede: Ein Forschungsüberblick. In: Jaki, S., Steiger, S. (eds) Digitale Hate Speech. J.B. Metzler, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-65964-9_8

Schwaiger, L. (2022). *Gegen die Öffentlichkeit: alternative Nachrichtenmedien im deutschsprachigen Raum*. [Against the Public: Alternative News Media in the German-speaking Area]. transcript Verlag.

Schwertberger, U., Rieger, D. (2021). Hass und seine vielen Gesichter: Eine sozial- und kommunikationswissenschaftliche Einordnung von Hate Speech. In: Wachs, S., Koch-Priewe, B., Zick, A. (eds) Hate Speech - Multidisziplinäre Analysen und Handlungsoptionen. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-31793-5_4

Shifman, L (2014) The cultural logic of photo-based meme genres. Journal of Visual Culture 13(3): 340–358.

Sponholz, L. (2020). Der Begriff "Hate Speech" in der deutschsprachigen Forschung: eine empirische Begriffsanalyse. [The Term 'Hate Speech' in German-Language Research: An Empirical Concept Analysis] SWS-Rundschau, 60(1), 43-65.

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.

Stahel, L., Weingartner, S., Lobinger, K. & Baier, D. (2022). Digitale Hassrede in der Schweiz: Ausmass und sozialstrukturelle Einflussfaktoren [Digital Hate Speech in Switzerland: Extent and Sociostructural Influencing Factors] 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).

Udupa, S. (2019). Nationalism in the digital age: Fun as a metapractice of extreme speech. *International Journal of Communication*, 3143-3163.

Vaught, S. E. (2012). Hate speech in a juvenile male prison school and in US schooling. Urban Review: Issues and Ideas in Public Education, 44(2), 239–264. https://doi.org/10.1007/s11256-011-0190-1

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

Wachs, S. & Wright, M. F. (2020). Associations between online hate victimization and perpetration: The buffering effects of technical and assertive coping. MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung, 16, 109–128. https://doi.org/10.21240/mpaed/jb16/2021.01.14.X

Walch, S. E., Ngamake, S. T., Francisco, J., Stitt, R. L., & Shingler, K. A. (2012). The Attitudes Toward Transgendered Individuals Scale: Psychometric Properties. *Archives of Sexual Behavior*, *41*(5), 1283-1291. https://doi.org/10.1007/s10508-012-9995-6

Wilson, R. A. & Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, *52*, 1029.

Danksagung

Bei der Umsetzung der Studie wurden wir von zahlreichen Personen und Organisationen unterstützt, denen wir unseren grössten Dank aussprechen möchten.

Nikki Böhler hat uns mit Know-how und grossem Netzwerk bei der Kommunikationskampagne zur Rekrutierung der Citizen Scientists unterstützt. Zahlreiche Organisationen haben auf unsere Studie und die Möglichkeit, Hassbilder für unser Forschungsprojekt spenden zu können, im Rahmen verschiedener Kommunikationsmassnahmen hingewiesen. Dazu zählen (in alphabetischer Reihenfolge): alliance F, Dachverband Schweizer Jugendparlamente, Fotomuseum Winterthur, Frauenzentrale Zürich, Frauenzentrale Kanton Glarus, #geschlechter gerechter, Info-Racisme Fribourg, Milchjugend, Museum für Kommunikation, OpenData.ch, Stiftung gegen Rassismus und Antisemitismus, WE/MEN, Zurich Pride & Zürich schaut hin. Ein besonderer Dank gebührt der Eidgenössischen Kommission gegen Rassismus (EKR) für die zur Verfügungstellung von Hassbildern, die auf ihrer Meldeplattform für rassistische Online-Hassrede eingingen

Grösster Dank gilt auch dem Dozierendenteam der Fachhochschule Graubünden Tanja Hess und Andreas Mädler sowie ihren Studierenden, die mit viel Kreativität und Können zahlreiche Counter-Hassbilder kreiert haben. Einige können auf dieser Webseite angesehen werden: https://www.mileva-institut.ch/counter-hassbilder.

Dankbar sind wir auch dem Bundesamt für Kommunikation (BAKOM), das durch die Förderung die Studien erst möglich machte.