universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Sprache, die polarisiert: Hate Speech und Counter Speech.

Eine experimentelle Untersuchung der Auswirkungen auf wahrgenommene soziale Identitätsbedrohung, stereotypbezogene Zuschreibungen und Interventionsbereitschaft

verfasst von | submitted by

Zara Weeke

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 841
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Publizistik- und Kommunikationswissenschaft
Betreut von | Supervisor:	Ass.-Prof. Mag. Dr. Kathrin Karsay Bakk.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Digitale Plattformen sind heute zentrale Orte öffentlicher Kommunikation, in denen gesellschaftliche Aushandlungsprozesse sichtbar und soziale Normen verhandelt werden (Sponholz, 2017, S. 20). Dort zeigen sich nicht nur vielfältige Formen politischer und sozialer Teilhabe, sondern auch Dynamiken der Abwertung, Polarisierung und Ausgrenzung. *Hate Speech* (HS, abwertende Sprache gegenüber sozialen Gruppen aufgrund ihrer Zugehörigkeit) ist in diesem Zusammenhang kein Randphänomen, sondern Ausdruck breiterer gesellschaftlicher Konflikte um Zugehörigkeit, Macht und die Grenzen akzeptabler öffentlicher Kommunikation. Laut einer Erhebung des Statistischen Bundesamtes hat im ersten Quartal 2025 etwa ein Drittel (34 %) der Internetnutzenden in Deutschland Beiträge im Zusammenhang mit Hassrede auf Webseiten oder in sozialen Medien wahrgenommen, was einen Anstieg gegenüber früheren Erhebungszeitpunkten darstellt. Rund 19,6 Millionen Personen im Alter von 16 bis 74 Jahren waren damit allein in Deutschland betroffen (Statistisches Bundesamt, 2025).

Empirische Analysen digitaler Plattformen verdeutlichen, dass sich diese Konflikte auch in veränderten Formen von Auseinandersetzungen widerspiegeln. Barth et al. (2023) beschreiben digitale Kommunikation als zunehmend geprägt von zugespitzten, konfliktgeladenen Auseinandersetzungen, Beleidigungen, Herabsetzungen und Formen aggressiver Kommunikation. Digitale Plattformen dienen damit sowohl dem Austausch zwischen unterschiedlichen gesellschaftlichen Gruppen als auch als Räume, in denen Konflikte sichtbarer und teilweise schärfer ausgetragen werden. Barth et al. (2023) fassen dieses Spannungsfeld als zentrale Ambivalenz digitaler Medienentwicklung (S. 210ff.).

Hate Speech ist dabei in gesellschaftliche Macht- und Ungleichheitsverhältnisse eingebettet. Digitale Gewalt und Abwertung stehen in engem Zusammenhang mit bestehenden Diskriminierungsstrukturen und sozialen Hierarchien (Barth et al., 2023, S. 212). Studien zeigen zudem, dass insbesondere Angehörige gesellschaftlich marginalisierter Gruppen überdurchschnittlich häufig Ziel von Hate Speech sind. Sie stellt damit nicht nur eine individuelle Belastung dar, sondern verweist auf strukturelle Ungleichheiten und gesellschaftliche Spannungen, die sich auf digitalen Plattformen verdichten (Schäfer et al., 2023, S. 554f.).

Die Folgen beschränken sich nicht auf direkt adressierte Personen. Forschung zeigt, dass auch Mitlesende hasserfüllter Kommunikation langfristig in ihren Einstellungen gegenüber bestimmten Gruppen beeinflusst werden können (You & Lee, 2019, S. 2ff.).

Wiederholte Konfrontation mit diskriminierenden Inhalten kann kognitive Aktivierungsprozesse und Desensibilisierung begünstigen (Hsueh et al., 2015, S. 588ff.; Schäfer et al., 2021, S. 4ff.). Für Betroffene selbst reichen die Auswirkungen häufig von sozialem Rückzug bis zu erheblichen psychischen Belastungen (Obermaier et al., 2014, S. 1492ff.). Beratungsstellen berichten zudem von Angst, Ohnmachtsgefühlen und Rückzugstendenzen infolge digitaler Hasskonfrontation (ZARA, 2025, S. 32ff.). Hate Speech betrifft damit nicht nur individuelle Erfahrungen, sondern auch Fragen gesellschaftlicher Teilhabe und Gleichberechtigung.

Vor diesem Hintergrund gewinnen rechtliche und plattformspezifische Rahmenbedingungen besondere Bedeutung. Sie strukturieren, welche Formen von Hate Speech sanktioniert werden und welche sichtbar bleiben. Plattformen wirken durch Community-Standards, Moderationspraktiken und algorithmische Funktionsweisen aktiv daran mit, kommunikative Normen auf digitalen Plattformen zu prägen (Barth et al., 2023, S. 211f.). Eine zentrale Herausforderung besteht darin, dass nicht alle schädlichen Inhalte strafrechtlich relevant sind. Der Bericht der Beratungsstelle GegenHassimNetz beschreibt dieses Spannungsfeld als „Lawful but Awful“ (ZARA, 2025, S.8), Inhalte, die rechtlich zulässig, aber dennoch verletzend, einschüchternd und gesellschaftlich problematisch sind. Solche Inhalte können bei Betroffenen Angst, Rückzug und Selbstzensur auslösen und so dazu beitragen, dass sich bestimmte Gruppen seltener an öffentlichen Debatten beteiligen (ZARA, 2025, S. 8f.).

Wie politisch und gesellschaftlich umkämpft Fragen der Moderation und Regulierung sind, zeigte sich besonders deutlich im Jahr 2025. Meta kündigte an, das Fact-Checking auf seinen Plattformen einzustellen und Richtlinien zu Hassrede und Missbrauch zu lockern, unter anderem in Bezug auf sexuelle Orientierung, Geschlechtsidentität und Einwanderungsstatus. CEO Mark Zuckerberg begründete dies mit dem Ziel, „restrictions on topics like immigration and gender that are out of touch with mainstream discourse“ zu entfernen (Ortutay, 2025). Auf der Meta Unternehmensseite hieß es zudem:

> Meta’s platforms are built to be places where people can express themselves freely. That can be messy. On platforms where billions of people can have a voice, all the good, bad and ugly is on display. But that’s free expression. (Meta, 2025)

Kritiker*innen befürchten, dass eine solche Neuausrichtung zu einer weiteren Zunahme von Hate Speech beitragen und präventive Moderationsmechanismen schwächen könnte (Ortutay, 2025).

Neben rechtlichen und plattformspezifischen Maßnahmen gewinnen kommunikative Gegenreaktionen zunehmend an Bedeutung. Als *Counter Speech* (CS) werden öffentlich sichtbare Reaktionen zusammengefasst, die Hate Speech widersprechen, Betroffene unterstützen und Grenzen akzeptabler Kommunikation deutlich machen (Garland et al., 2022, S. 1ff.). Auch Beratungsstellen betonen die Rolle digitaler Zivilcourage, bei der nicht nur Betroffene, sondern auch alle Mitlesenden Verantwortung übernehmen (ZARA, 2025, S. 30f.).

1.1 Problemstellung und Relevanz des Themas

Kommentarbereiche sozialer Medien sind besonders relevante Untersuchungsräume, da sich gesellschaftliche Aushandlungsprozesse hier in verdichteter Form beobachten lassen. Als öffentlich einsehbare Interaktionsräume tragen sie zur (Re-)Produktion sozialer Identitäten, Zugehörigkeitsordnungen und Machtverhältnisse bei. Gleichzeitig prägen sie Vorstellungen darüber, welche Formen der Kommunikation als akzeptabel gelten.

Empirische Studien zeigen, dass wiederholte Konfrontation mit Hate Speech negative Stereotype verstärken, Distanz fördern und Wahrnehmungen sozialer Normen verändern kann. Insbesondere Angehörige gesellschaftlich marginalisierter Gruppen erleben Hate Speech als Bedrohung kollektiver Zugehörigkeit (Hsueh et al., 2015, S. 588ff.; Schäfer et al., 2021, S. 4ff.; Soral et al., 2017, S. 1ff.).

Gleichzeitig begegnen Mitlesende Hate Speech in der Regel nicht isoliert, sondern im Kontext weiterer Kommentare, Reaktionen und Bewertungen. Zustimmung, Widerspruch oder ausbleibende Reaktionen anderer können als Hinweise auf geltende soziale Normen gedeutet werden.

Vor diesem Hintergrund stellt sich die Frage nach der Wirkung von Counter Speech. Sie kann Hasskommentare als normabweichend markieren, Solidarität mit Betroffenen ausdrücken und alternative Deutungen anbieten. Ob und unter welchen Bedingungen Counter Speech tatsächlich Einstellungen, emotionale Reaktionen und Verhaltensintentionen beeinflusst, ist bislang nicht eindeutig geklärt. Die bisherigen Befunde sind uneinheitlich und deuten darauf hin, dass mögliche Effekte stark vom Kontext der konkreten Ausgestaltung der Counter Speech sowie von den Voreinstellungen der Mitlesenden abhängen (Schäfer et al., 2023, S. 570ff.).

Bislang ist zudem nur unzureichend untersucht, wie Mitlesende in realitätsnahen Kommentarbereichen auf Hate Speech reagieren, wenn gleichzeitig Counter Speech sichtbar ist. Unklar ist insbesondere, welche affektiven, kognitiven und handlungsbezogenen

Reaktionen dabei auftreten und inwiefern Counter Speech diese Reaktionen beeinflussen kann.

1.2 Zielsetzung und Forschungsfrage

Vor diesem Hintergrund zielt die vorliegende Studie darauf ab, die Wirkungen von Hate Speech und Counter Speech in digitalen Kommentaröffentlichkeiten differenziert zu untersuchen. Im Mittelpunkt steht ein experimenteller Vergleich von Hate Speech ohne Counter Speech und Hate Speech mit Counter Speech in einer realitätsnah gestalteten Instagram-Kommentarsituation. Analysiert werden drei Wirkungsebenen: wahrgenommene soziale Identitätsbedrohung als affektiv-gruppenbezogene Reaktion, stereotypbezogene Zuschreibungen als kognitive Reaktion sowie Interventionsbereitschaft als handlungsbezogene Reaktion. Zusätzlich wird variiert, ob die Counter Speech aus der Mehrheitsgesellschaft (CS-M) oder aus der ethnischen Minderheit (CS-eM) stammt (CS-Quelle). Es wird auch geprüft, ob sich die Wirkungen zwischen Teilnehmenden aus der Mehrheitsgesellschaft (TG-M) und ethnischen Minderheiten (TG-eM) unterscheiden. Daraus ergibt sich folgende Forschungsfrage:

> Wie beeinflusst die Exposition gegenüber Hate Speech im Vergleich zu Hate Speech mit Counter Speech (1) die wahrgenommene soziale Identitätsbedrohung, (2) stereotypbezogene Zuschreibungen sowie (3) Interventionsbereitschaft, und unterscheiden sich diese Wirkungen in Abhängigkeit von (a) der CS-Quelle (CS-M vs. CS-eM) sowie (b) der Teilnehmendengruppen (TG-M vs. TG-eM)?

1.3 Aufbau der Studie

Das nächste Kapitel entwickelt den theoretischen Rahmen zu Hate Speech, Counter Speech sowie den zugrunde liegenden Wirkmechanismen. Darauf aufbauend fasst Kapitel 3 den empirischen Forschungsstand zusammen und zeigt zentrale Forschungslücken auf. Kapitel 4 führt in die vorliegende Studie ein und leitet die Forschungsfrage, die Hypothesen sowie eine explorative Forschungsfrage her. Anschließend beschreibt Kapitel 5 das methodische Vorgehen, während in Kapitel 6 die Ergebnisse der Studie präsentiert werden. Diese werden in Kapitel 7 angesichts der theoretischen Ansätze und des bisherigen Forschungsstands diskutiert. Abschließend fasst Kapitel 8 die zentralen Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Forschung.

2 Theoretischer Rahmen

Dieses Kapitel bestimmt Hate Speech und Counter Speech als kommunikative Phänomene in digitalen Kommentaröffentlichkeiten und die theoretischen Mechanismen durch die beide wirksam werden können. Dazu werden zunächst zentrale Arbeitsdefinitionen eingeführt. Darunter Hate Speech*, ethnische Minderheiten* und *Mehrheitsgesellschaft* sowie Counter Speech. Darauf aufbauend werden jene theoretischen Perspektiven dargestellt, die zur Erklärung der untersuchten Wirkungen herangezogen werden. Dazu zählen Ansätze sozialer Identitätsbedrohung zur Erklärung affektiver-gruppenbezogener Reaktionen, Priming-Ansätze zur Darlegung kurzfristiger kognitiver Aktivierungen sowie das Bystander-Modell zur Strukturierung handlungsbezogener Interventionsprozesse. Auf dieser Basis wird ein konzeptioneller Rahmen entwickelt, der die theoretische Grundlage für die in Kapitel 4 formulierten Hypothesen bildet.

2.1 Hate Speech

Hate Speech (HS) wird in der Forschung allgemein als kommunikative Handlung verstanden, die Personen aufgrund ihrer Zugehörigkeit zu sozialen Gruppen abwertet, entmenschlicht oder bedroht (Esau, 2023, S. 453; Castaño-Pulgarín et al., 2021, S. 5; Ruscher, 2024, S. 2). Dabei handelt es sich nicht zwingend um rein verbale Kommunikation, sondern auch um visuelle oder symbolische Ausdrucksformen, etwa Bilder, Memes oder Zeichen, die gruppenbezogene Abwertung transportieren (Schäfer et al., 2023, S. 554). Trotz intensiver Forschung existiert bis heute keine disziplinübergreifend einheitliche Definition (Kansok-Dusche et al., 2022, S. 1, 10; Paasch-Colberg et al., 2021, S. 172). Brown (2017) beschreibt Hate Speech daher als *family resemblance*-Konzept. Einzeldefinitionen überlappen sich in wichtigen Aspekten wie Gruppenbezug und abwertende Intention, setzen allerdings unterschiedliche Schwerpunkte (S. 593–604).

Eine verbreitete Systematisierung bietet Sellars (2016), der wiederkehrende Merkmalsdimensionen wie Gruppenbezug, abwertenden Ausdruck, Kontextbedingungen und potenzielle Schädigung hervorhebt (S. 24–31). Aufbauend darauf identifizieren Kansok-Dusche et al. (2022) vier zentrale Strukturmerkmale. Abwertende Ausdrucksformen, Bezug auf Gruppenmerkmale, potenzielle oder tatsächliche Schädigung auf individueller oder kollektiver Ebene sowie Wirkung, die Ungleichheit legitimiert (S. 11).

Ergänzend betont Gelber (2019), dass Hate Speech in bestehende Diskriminierungs- und Ungleichheitsverhältnisse eingebettet ist und sich häufig gegen gesellschaftlich ohnehin benachteiligte Gruppen richtet (S. 407ff.). Diese machtkritische Perspektive ist für die

vorliegende Studie zentral, da sie Hate Speech nicht als isolierte Beleidigung, sondern als Ausdruck struktureller Ungleichheit konzeptualisiert.

Zudem ist Hate Speech von verwandten Formen problematischer Kommunikation abzugrenzen. Cyberbullying beschreibt überwiegend interpersonale, meist wiederholte Angriffe zwischen Individuen und zielt nicht zwingend auf soziale Gruppenzugehörigkeit ab. Davon zu unterscheiden ist Cyberhate, ein Oberbegriff für gruppenbezogene Feindseligkeit im digitalen Raum, der unterschiedliche Ausdrucksformen digitaler Abwertung umfasst. Für Satire oder Ironie ist eine Einordnung als Hate Speech hingegen nicht ausgeschlossen, sofern Gruppenbezug, abwertender Ausdruck und Schädigungspotenzial vorliegen (Kansok-Dusche et al., 2022, S. 9ff.).

Auf Grundlage der genannten Literatur wird Hate Speech in der vorliegenden Studie wie folgt definiert: Hate Speech bezeichnet kommunikative Ausdrucksformen, die Personen aufgrund tatsächlicher oder zugeschriebener Gruppenmerkmale systematisch abwerten, entmenschlichen oder Gewalt legitimieren und dadurch bestehende Machtungleichheiten reproduzieren oder die gesellschaftliche Teilhabe der Betroffenen beeinträchtigen.

Diese Definition stellt sicher, dass Hate Speech nicht auf sprachliche Aggressivität begrenzt wird, sondern auch den gruppenbezogenen Abwertungsgehalt sowie das implizierte Schädigungspotenzial berücksichtigt.

2.2 Ethnische Minderheiten und Mehrheitsgesellschaft

Da Hate Speech auf kollektive Identitäten zielt und insbesondere gesellschaftlich benachteiligte Gruppen betrifft, ist eine präzise begriffliche Klärung der in dieser Studie untersuchten Zielgruppen erforderlich.

In Anlehnung an soziologische Konzepte ethnischer Ungleichheit werden ethnische Minderheiten als soziale Gruppen verstanden. Sie werden über zugeschriebene oder selbst beanspruchte ethnische Zugehörigkeit (z. B. Abstammung, Sprache, Traditionen) kategorisiert und sind kontextspezifisch in Macht-, Ressourcen- und Deutungsstrukturen strukturell benachteiligt. Entscheidend ist dabei nicht die numerische Größe der Gruppe, sondern ihre gesellschaftliche Position innerhalb bestehender Machtasymmetrien (Dunn, 2021, Kapitel 2.1).

Als Mehrheitsgesellschaft werden entsprechend jene sozialen Gruppen gefasst, die gesellschaftliche Normen, kategoriale Grenzziehungen und institutionelle Rahmenbedingungen prägen sowie über überproportionalen Zugang zu ökonomischen, politischen und symbolischen Ressourcen verfügen (Dunn, 2021, Kapitel 2.1).

Begrifflich ist ethnisch im Sinne kulturell-historischer Herkunftsbezüge klar von national als rechtlicher Kategorie der Staatsangehörigkeit sowie von kulturell als Ausdruck geteilter Praktiken oder Werte zu unterscheiden (Sökefeld, 2007, S. 32, 47). Für diese Studie ist diese Differenzierung zentral, da Hate Speech häufig ethnische Minderheiten adressiert, unabhängig von Staatsangehörigkeit oder tatsächlicher kultureller Praxis.

Auf dieser Grundlage werden die zentralen Begriffe für die vorliegende Studie wie folgt festgelegt: Ethnische Minderheiten bezeichnen soziale Gruppen, die über (zugeschriebene oder selbst beanspruchte) ethnische Herkunftsmerkmale definiert werden und in gesellschaftlichen Macht-, Ressourcen- und Deutungsstrukturen eine benachteiligte Position einnehmen. Als Mehrheitsgesellschaft gelten jene Gruppen, die gesellschaftliche Normen prägen und über dominanten Status sowie institutionelle, ökonomische und symbolische Ressourcen verfügen.

Die Unterscheidung zwischen ethnischen Minderheiten und Mehrheitsgesellschaft bildet die Grundlage für gruppenvergleichende Analysen dieser Studie. Sie ist insbesondere deshalb theoretisch bedeutsam, da die zugrunde liegenden Wirkmechanismen von Hate Speech und Counter Speech je nach Gruppenzugehörigkeit unterschiedlich ausfallen können (Dunn, 2021, Kapitel 2.4).

2.3 Counter Speech

Counter Speech (CS) wird in der Forschung als zentrale Gegenreaktion auf Hate Speech verstanden, bei der Nutzer*innen öffentlich sichtbar widersprechen, Grenzen markieren und Unterstützung für angegriffene Gruppen signalisieren und somit Zivilcourage zeigen. Im Unterschied zu plattformseitigen Eingriffen, wie dem Löschen oder Sperren von Inhalten, zielt Counter Speech nicht primär auf die Entfernung dieser Äußerungen, sondern auf deren soziale Einordnung und Aushandlung innerhalb öffentlicher Kommunikationsräume ab (Garland et al., 2022, S. 1ff.; Obermaier et al., 2021, S. 2ff.; Ruscher, 2024, S. 48ff.).

Die Forschung unterscheidet verschiedene Stilformen von Counter Speech. Bartlett & Krasodomski-Jones (2015) differenzieren zwischen faktenbasierten Ansätzen, die auf sachliche Korrekturen abzielen, und solchen, die selbst aggressive oder herabsetzende Elemente enthalten können (S. 11). Daran anknüpfend beschreiben neuere Arbeiten mehrere wiederkehrende Stiltypen. Dazu gehören faktenbasierte Richtigstellungen, empathisch-unterstützende Beiträge, humorvolle Reaktionen sowie normativ-appellative Formen, die auf soziale Normen, Werte oder Regeln verweisen (Obermaier et al., 2021, S. 2ff.). Neben der sprachlichen Umsetzung könnte für die Wirkung von Counter Speech auch entscheidend sein, von wem sie geäußert wird. Die wahrgenommene CS-Quelle kann als Hinweis darauf

fungieren, ob eine Reaktion als Hinweis gruppenspezifischer Betroffenheit (CS aus der ethnischen Minderheit) oder als gruppenübergreifende Solidarität (CS aus der Mehrheitsgesellschaft) interpretiert wird. Effekte der CS-Quelle werden theoretisch insbesondere über soziale Identität und Normsignale erklärt und als mögliche Moderatoren der Wirkung von Counter Speech diskutiert (Obermaier et al., 2021, S. 2ff.; Ruscher, 2024, S. 48ff.).

Vor diesem Hintergrund wird Counter Speech in der vorliegenden Studie wie folgt definiert: Counter Speech sind öffentlich sichtbare Antworten von Nutzer*innen auf Hate Speech, die dem Inhalt explizit widersprechen und/oder die angegriffene Gruppe unterstützen. Dabei können unter anderem faktenbasierte Hinweise, solidarische Unterstützung, humorvolle Entschärfungen oder normative Appelle eingesetzt werden.

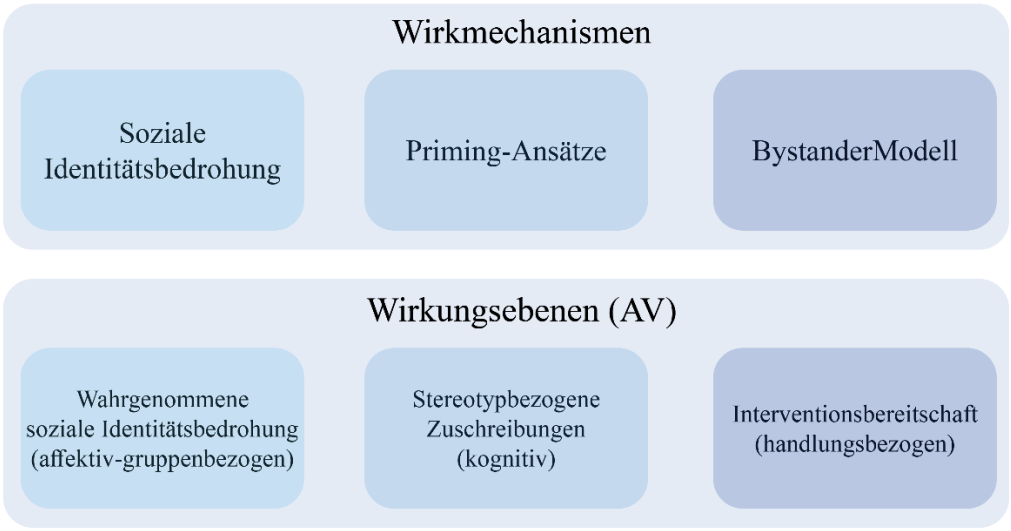
Für die vorliegende Studie ist Counter Speech zentral, weil sie in digitalen Kommentarbereichen eine sichtbare Gegenposition zu Hate Speech darstellt. Sie kann Hate Speech als normabweichend markieren und zugleich soziale Unterstützung für die angegriffene Gruppe signalisieren.

2.4 Wirkmechanismen von Hate Speech und Counter Speech

Um Wirkungen auf Mitlesende von Hate Speech und Counter Speech empirisch untersuchen zu können, bedarf es theoretischer Modelle, die erklären, wie abwertende Inhalte wahrgenommen, kognitiv verarbeitet und in Handlungsabsichten übersetzt werden. Die vorliegende Studie integriert drei theoretische Perspektiven, die diese Wirkmechanismen erklären. Wahrgenommene soziale Identitätsbedrohung zur Erklärung affektiv-gruppenbezogener Reaktionen, Priming-Ansätze zur Erklärung kognitiver Aktivierungen sowie das Bystander-Modell zur Erklärung handlungsbezogener Prozesse der Verantwortungsübernahme und Interventionsbereitschaft. Die Zusammenführung dieser Perspektiven ist in Abbildung 1 schematisch dargestellt.

Abbildung 1

Wirkmechanismen und -ebenen von Hate Speech und Counter Speech



Diese Perspektiven werden im Folgenden näher ausgeführt und bilden gemeinsam den theoretischen Rahmen der vorliegenden Studie.

2.4.1 Soziale Identitätsbedrohung

Die Soziale-Identitäts-Theorie (SIT) nach Tajfel und Turner (1979) beschreibt, dass soziale Gruppenzugehörigkeiten einen zentralen Bestandteil des Selbstkonzepts bilden. Individuen kategorisieren sich selbst und andere in soziale Gruppen, identifizieren sich mit relevanten Ingroup-Kategorien und bewerten diese im sozialen Vergleich, um eine positive soziale Identität aufrechtzuerhalten (Tajfel & Turner, 1979, S. 34, 41f.). Diese Prozesse sind häufig mit intergruppalen Effekten wie Ingroup-Favorisierung, stereotypbezogener Zuschreibung und dem Streben nach Abgrenzung verbunden (Brown, 2000, S. 750ff.; Wang et al., 2021, S. 461).

Soziale Identitätsbedrohung bezeichnet Situationen, in denen Personen ihre Zugehörigkeit zu einer sozialen Gruppe als abgewertet oder bedroht wahrnehmen. Eine solche Bedrohung kann bereits dadurch entstehen, dass negative gesellschaftliche Stereotype über die eigene Gruppe bekannt sind, unabhängig davon, ob ihnen aktiv zugestimmt wird (Steele, 1997, S. 614f.). Empirisch ist soziale Identitätsbedrohung mit Stressreaktionen wie erhöhter Wachsamkeit bzw. verstärkter Aufmerksamkeit für Bedrohungssignale, Angst oder verringerter kognitiver Leistungsfähigkeit verbunden und kann Bewältigungsstrategien wie Rückzug, Distanzierung, verstärkte Gruppenidentifikation oder kollektives Handeln auslösen (Brown, 2000, S. 750ff.; Major & O'Brien, 2004, S. 795ff.).

Studien zeigen, dass Hate Speech Wirkungen wie soziale Identitätsbedrohung auslösen kann (Obermaier et al., 2021, S.1), und besonders schwere Formen von Hate Speech zu

Vermeidungsverhalten oder Rückzug aus Intergruppeninteraktionen führen können (Van Houtven et al., 2024, S. 934).

Für die vorliegende Studie ist wahrgenommene soziale Identitätsbedrohung von zentraler Bedeutung, da Hate Speech in digitalen Kommentaröffentlichkeiten als öffentlich sichtbare Abwertung kollektiver Zugehörigkeit verarbeitet wird und damit affektiv-gruppenbezogene Reaktionen auslösen kann. Sichtbare Counter Speech kann diese Situation verändern, indem sie die Abwertung explizit zurückweist und Unterstützung für die angegriffene Gruppe signalisiert, wodurch die Wahrnehmung sozialer Identitätsbedrohung potenziell abgeschwächt wird (Obermaier et al., 2021, S. 5ff.; Schieb & Preuss, 2016, S. 19).

Vor diesem Hintergrund unterscheidet die Studie zwischen der CS-Quelle (Mehrheitsgesellschaft (CS-M) vs. ethnischen Minderheit (CS-eM)) sowie der Teilnehmendengruppen (Mehrheitsgesellschaft (TG-M) und ethnische Minderheiten (TG-eM)).

2.4.2 Priming-Ansätze

Zur Erklärung kurzfristiger kognitiver Effekte ist Priming zentral. Priming basiert auf dem assoziativen Netzwerkmodell von Collins und Loftus (1975), demzufolge Wissenselemente (z. B. Begriffe, Bewertungen, Stereotype) als miteinander verknüpfte Knoten im Gedächtnis repräsentiert sind. Wird ein Knoten aktiviert, erhöht sich durch sogenannte *Spreading Activation* die Zugänglichkeit semantisch verwandter Inhalte (Collins & Loftus, 1975, S. 407–428; Schemer, 2013, S. 154f.).

Higgins (1996) erläutert, dass Priming-Effekte vor allem dann eintreten, wenn ein aktiviertes Konzept sowohl verfügbar, das bedeutet stark verankert oder häufig aktiviert, als auch situativ anwendbar auf die gegenwärtige Wahrnehmungs- oder Urteilssituation ist. Medienpriming beschreibt den Prozess, durch den mediale Inhalte kurzfristig die Zugänglichkeit bestimmter Wissensbestände erhöhen und dadurch nachfolgende Urteile beeinflussen können, ohne dass dieser Einfluss bewusst reflektiert werden muss. Solche Effekte können auch nach der unmittelbaren Rezeption fortwirken (Higgins, 1996, S.147f.; Peter, 2002, S. 20; Power et al., 1996, S. 40).

Im Unterschied zu Agenda-Setting oder Framing steht beim Priming nicht die Gewichtung von Themen oder die Interpretation von Inhalten im Vordergrund. Vielmehr geht es um die Aktivierung von Bewertungskriterien, die in anschließenden Urteilen mit erhöhter Wahrscheinlichkeit herangezogen werden (Peter, 2002, 24f.). Power et al. (1996) fassen diesen Mechanismus prägnant zusammen: „the activation of one category or schema—for

example, a cultural stereotype—increases the likelihood that the same category will be used in subsequent judgments“ (S.37).

Empirische Forschung zeigt, dass Hate Speech einen besonders starken Stimulus für stereotypbezogene Zuschreibungen darstellt. Durch explizite gruppenbezogene Zuschreibungen kann Hate Speech sowohl offensichtliche als auch implizite Stereotype aktivieren (Hsueh et al., 2015, S. 570f.). Zudem kann wiederholte Exposition Prozesse wie Desensibilisierung und Normverschiebungen begünstigen und Distanz gegenüber den genannten Gruppen erhöhen (Soral et al., 2017, 1ff.; Schäfer et al., 2023, S. 554ff.).

Counter Speech wird in diesem Zusammenhang als potenzieller kognitiver und normativer Gegenimpuls diskutiert. Theoretisch kann Counter Speech stereotype Aktivierungen abschwächen, indem sie alternative Deutungsangebote bereitstellt oder die Anwendbarkeit aktivierter Stereotype durch normative Gegensignale infrage stellt (Schieb & Preuss, 2016, S. 19; Obermaier et al., 2021, S. 5ff.).

Gleichzeitig zeigen Untersuchungen, dass Aktivierungen aufgrund von Priming nicht vollständig aufgehoben werden können. Stereotype Inhalte bleiben weiterhin präsent, während gegenstereotype Informationen oft weniger zugänglich sind als stereotype (Schäfer et al., 2023, S. 560; Power et al., 1996, S. 40ff.).

Vorbestehende Einstellungen können beeinflussen, in welchem Ausmaß Hate Speech stereotype Assoziationen aktiviert (Schäfer et al., 2023, S. 554ff.). Auch Peter (2002) weist darauf hin, dass Stärke und Dauer von Priming-Effekten zwischen Personen und Situationen erheblich variieren können (Peter, 2002, S. 22).

Für die vorliegende Studie folgt daraus die Erwartung, dass Hate Speech kurzfristig stereotypbezogene Zuschreibungen beeinflussen und Counter Speech als konkurrierendes Norm- und Deutungssignal diese Effekte potenziell reduzieren kann.

2.4.3 Bystander-Modell

Während Ansätze sozialer Identitätsbedrohung und Priming erklären, wie Hate Speech wahrgenommen sowie kognitiv und affektiv verarbeitet wird, erfordert die Analyse handlungsbezogener Reaktionen ein Modell, das erklärt, unter welchen Bedingungen Mitlesende intervenieren. Ein zentraler theoretischer Bezugspunkt ist das Bystander-Modell von Latané und Darley (1970), das Hilfeverhalten als mehrstufigen Entscheidungsprozess beschreibt. Eine Situation muss wahrgenommen, als Notfall interpretiert, mit persönlicher Verantwortung beantwortet, mit Handlungsoptionen verknüpft und schließlich umgesetzt werden. Zentral ist dabei die Übernahme persönlicher Verantwortung, da ohne sie Intervention unwahrscheinlich bleibt. Empirische Forschung zeigt zudem, dass die

Anwesenheit weiterer Beobachtender diese Verantwortungsübernahme erschweren kann (Fischer et al., 2011, S. 518f.; Obermaier et al., 2014, S. 1493f.). Dieser Effekt äußert sich darin, dass Verantwortung auf andere verteilt wird und Einzelne sich weniger verpflichtet fühlen zu handeln (Fischer et al., 2011, S. 518f.; Obermaier et al., 2014, S. 1493f.).

Ergänzend identifiziert die Literatur weitere hemmende Prozesse, darunter *diffusion of responsibility*, *audience inhibition*, also die Sorge vor negativer sozialer Bewertung, sowie Unsicherheiten in der Situationsbewertung, die entstehen, wenn andere Beobachtende nicht reagieren. Das Ausbleiben sichtbarer Interventionen kann dazu führen, dass Nicht-Eingreifen als soziale Norm interpretiert wird (You & Lee, 2019, S. 2).

Zugleich weisen You und Lee (2019) darauf hin, dass Bystander-Verhalten kein binärer Zustand ist. Je nach sozialer Situation und Risikowahrnehmung unterscheiden sie zwischen inaktivem Verhalten, passivem Beobachten, helfender Reaktion (z. B. Meldung an Plattformen, Unterstützung der Betroffenen Person) oder konfrontativer Intervention. Welche Form gewählt wird, hängt unter anderem von situativen Einschätzungen, Empathie, Selbstwirksamkeit und wahrgenommenem Risiko ab (You & Lee, 2019, S. 2f.).

Mehrere Studien übertragen diese Prozesse auf digitale Kommunikationskontexte. Die fehlende physische Anwesenheit anderer, die unklare Publikumsgröße und eingeschränkte soziale Rückmeldungen, können Verantwortungsübernahme reduzieren, gleichzeitig aber auch Intervention erleichtern, etwa durch niedrigschwellige Handlungsmöglichkeiten oder geringere unmittelbare soziale Sanktionen (Obermaier et al., 2014, S. 1493f.; Obermaier et al., 2023, S. 817ff.; You & Lee, 2019, S. 2f.).

Für Hate-Speech-Kontexte ergibt sich daraus eine ambivalente Rolle für Counter Speech. Einerseits kann Counter Speech als Normsignal wirken, das die Situation eindeutig als grenzüberschreitend markiert und die Interventionsbereitschaft erhöht. Andererseits kann sie den Eindruck erzeugen, dass bereits gehandelt wurde, wodurch erneute Verantwortungsübernahme sinkt (Obermaier et al., 2014, S. 1493ff.). Damit bleibt theoretisch zu klären, ob Counter Speech im jeweiligen Kontext eher mobilisierend oder entlastend wirkt, ein zentraler Grund dafür, warum dieser Mechanismus in dieser Studie explorativ geprüft wird (siehe Kapitel 4).

2.4.4 Zusammenfassung und Relevanz für die vorliegende Studie

Die dargestellten Ansätze bilden gemeinsam den theoretischen Rahmen der vorliegenden Studie. Ansätze sozialer Identitätsbedrohung erklären, warum gruppenbezogene Abwertung in Hate Speech als Bedrohung der sozialen Identität erlebt werden kann und warum diese Reaktion je nach Gruppenzugehörigkeit unterschiedlich ausfallen dürfte. Priming-Ansätze

erklären, warum Hate Speech stereotypbezogene Zuschreibungen kurzfristig zugänglicher macht und warum Counter Speech als konkurrierendes Deutungs- und Normsignal potenziell dämpfend wirkt, ohne primingbasierte Aktivierungen notwendigerweise vollständig aufzuheben. Das Bystander-Modell strukturiert schließlich, wie Mitlesende Situationen bewerten und Verantwortung übernehmen und warum Counter Speech sowohl handlungsaktivierend als auch interventionshemmend sein kann.

Kapitel 3 greift diese theoretischen Annahmen auf, ordnet sie in den bestehenden Forschungsstand ein und stellt zentrale empirische Befunde zu Hate Speech und Counter Speech vergleichend dar. Auf diese Weise werden die Befunde in ihren jeweiligen Forschungskontext eingebettet und ihre Bedeutung für die vorliegende Untersuchung präzisiert.

3 Empirischer Forschungsstand zu Hate Speech und Counter Speech

Dieses Kapitel bündelt zentrale empirische Befunde und ordnet sie entlang der dort eingeführten Wirkmechanismen ein. Es zeigt konsistente Befundmuster ebenso wie kontextabhängige Ergebnisse und leitet daraus methodische und inhaltliche Forschungslücken ab, an die die vorliegende Studie anschließt.

Zunächst werden einschlägige Überblicks- und Reviewarbeiten sowie zentrale Problemstellungen des Forschungsfeldes zusammengefasst. Anschließend werden dominante methodische Zugänge und wiederkehrende Forschungsdesigns der empirischen Hate-Speech-Forschung systematisch dargestellt. Darauf aufbauend werden typische Ausprägungen von Hate Speech, relevante Zielgruppen und Plattformkontexte, empirisch belegte Wirkungen sowie zentrale Gegenmaßnahmen mit besonderem Fokus auf Counter Speech diskutiert. Abschließend werden die dargestellten Befunde in einer kritischen Synthese zusammengeführt, aus der zentrale Forschungslücken abgeleitet werden.

3.1 Überblick über den Forschungsstand

Mit der Verbreitung von sozialen Medien hat sich die Forschung von Hate Speech deutlich ausgeweitet und umfasst heute verschiedene disziplinäre Perspektiven, wie rechtliche, kommunikationswissenschaftliche, sozialpsychologische und informatikbezogene Ansätze.

Überblicks- und Reviewarbeiten zeigen, dass die empirische Hate-Speech-Forschung stark gewachsen ist, dennoch in Bezug auf ihren Untersuchungsdesigns, Operationalisierungen und betrachteten Kontexten erheblich variiert (Castaño-Pulgarín et al., 2021, S. 2; Barth et al., 2023, S. 213f.; Waqas et al., 2019, S. 2; Paz et al. 2020, S. 1ff.). Reviews verdeutlichen, dass empirische Studien sehr unterschiedliche Erscheinungsformen gruppenbezogener Abwertung untersuchen, was die Vergleichbarkeit von Befunden erschwert (Bührer et al., 2024, 1ff.; Kansok-Dusche et al., 2022, S. 11; Khaleghipour et al., 2025, S. 2). Barth et al. (2023) beschreiben diese Situation als „semantic blurriness and polysemy“ (S. 213), da zentrale Begriffe unterschiedlich verwendet werden und direkte Vergleiche zwischen Befunden dadurch erschwert werden.

Für die Einordnung der folgenden Befunde ist daher zentral, welche Designs, Kontexte und Operationalisierungen den jeweiligen Studien zugrunde liegen. Entsprechend werden im folgenden Abschnitt zentrale methodische Zugänge und wiederkehrende Designmuster der empirischen Hate-Speech-Forschung dargestellt.

3.2 Methodische Zugänge und Forschungsdesign

Überblicksarbeiten zeigen, dass die empirische Hate-Speech-Forschung methodisch vielfältig ist, gleichzeitig aber klare Schwerpunkte aufweist. Quantitative Studien dominieren, insbesondere querschnittliche Befragungsdesigns, die Häufigkeiten, Korrelationen und Selbstauskünfte zu Wahrnehmungen, Einstellungen und Verhaltensintentionen erfassen (Barth et al., 2023, S. 210ff.; Bührer et al., 2024, S. 8ff.; Castaño-Pulgarín et al., 2021, S. 2ff.). Waqas et al. (2019) fassen dies zusammen, indem sie betonen, dass „online hate is a complex phenomenon—with its definition depending on theoretical paradigms, disciplines, and forms of victimization“ (S. 2).

Bührer et al. (2024) halten entsprechend fest, dass „most studies employed self-report cross-sectional surveys“ (S. 8). Solche Designs ermöglichen eine breite Erfassung subjektiver Wahrnehmungen, bilden allerdings vor allem Momentaufnahmen ab und sind nur eingeschränkt geeignet, kausale Wirkmechanismen zu prüfen. Längsschnittliche und experimentelle Designs sind zwar vorhanden, jedoch deutlich weniger verbreitet (Castaño-Pulgarín et al., 2021, S. 2ff.).

Ferner weisen Überblicksarbeiten auf systematische Stichprobenverzerrungen hin. Viele Studien konzentrieren sich auf einzelne Plattformen und verwenden hauptsächlich westlich geprägte Stichproben (WEIRD-Bias). Dadurch wird die vergleichende Untersuchung plattformspezifischer Kommunikationslogiken und Zielgruppenunterschiede nur begrenzt vorgenommen (Bührer et al., 2024, S. 8ff.).

Die Aussagekraft der Ergebnisse hängt stark von den spezifischen Forschungsdesigns, Stichproben und Plattformkontexten ab. Der nächste Abschnitt bündelt zentrale empirische Systematisierungen typischer Formen von Hate Speech.

3.3 Formen und Ausprägungen von Hate Speech

Aufbauend auf den in Kapitel 2.1 entwickelten begrifflichen Grundlagen richtet dieser Abschnitt den Fokus auf die empirische Operationalisierung von Hate Speech in der bisherigen Forschung. Im Mittelpunkt steht dabei die Frage, entlang welcher wiederkehrenden Inhaltsdimensionen und Ausdrucksformen Studien entsprechende Beiträge systematisch erfassen.

Empirische Arbeiten strukturieren Hate Speech häufig entlang stabiler inhaltlicher Kategorien, die sich über unterschiedliche Kontexte hinweg beobachten lassen. Dazu zählen insbesondere pauschalisierende negative Zuschreibungen gegenüber sozialen Gruppen, entmenschlichende Darstellungen durch Tier-, Krankheits- oder Objektmetaphern sowie explizite oder implizite Gewaltbezüge, etwa in Form von Gewaltlegitimierung oder

Gewaltanndrohung. Solche Kategorien ermöglichen es, abwertende Kommunikationsinhalte systematisch zu erfassen und zwischen Studien vergleichbar zu machen. Zusätzlich berücksichtigen Studien die konkreten Ausdrucksformen, in denen diese Inhalte realisiert werden. Dazu gehören sprachliche Ausdrucksformen wie beleidigende Bemerkungen über Gruppen oder pauschale Verallgemeinerungen genauso wie visuelle oder symbolische Darstellungen, wie zum Beispiel Memes, Bilder oder Embleme, die eine negative Haltung gegenüber Gruppen zum Ausdruck bringen (Paasch-Colberg et al., 2021, S. 175ff.; Ruscher, 2024, S. 6ff.).

Diese Systematisierungen erleichtern es, realitätsnahe Kommunikationsbeispiele theoriegeleitet auszuwählen und zugleich kontrolliert zu variieren. Sie bilden damit eine wichtige Grundlage für experimentelle Untersuchungen, in denen einzelne Merkmale von Hate Speech gezielt verändert werden können.

3.4 Adressierte Gruppen und Plattformkontexte

Empirische Untersuchungen belegen, dass die Sichtbarkeit, Verbreitung und Einordnung von Hate Speech stark durch den sozialen Kontext und die spezifischen Plattformstrukturen beeinflusst werden. Dabei richten sich Hate Speech besonders häufig gegen Gruppen, die in gesellschaftlichen Macht- und Ressourcenstrukturen benachteiligt sind (Schäfer et al., 2023, S. 3; Schneiders, 2021, S. 276; Sponholz, 2017, 59ff.). Adressiert werden typischerweise Gruppen„on the basis of characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation" (Castaño-Pulgarín et al., 2021, S. 1).

Forschungsergebnisse zeigen, dass Hate Speech nicht nur direkt Betroffene belastet, sondern auch indirekte Wirkungen entfaltet. Insbesondere marginalisierte Gruppen erleben *vicarious victimization* (stellvertretende Viktimisierung), wenn Abwertung gegenüber Mitgliedern der eigenen sozialen Gruppe sichtbar wird (Khaleghipour et al., 2025, S. 7f.).

Die Sichtbarkeit und Reichweite von Hate Speech werden zudem maßgeblich durch plattformspezifische Gegebenheiten geprägt. Digitale Plattformen fungieren nicht nur als neutrale Kommunikationsräume, sondern strukturieren durch algorithmische Auswahl, Interaktionslogiken und technische Rahmenbedingungen, welche Inhalte wahrgenommen und verstärkt werden (Cinelli et al., 2021, S. 5f.; Sponholz, 2017, S. 429ff.). Studien verweisen darauf, dass insbesondere algorithmische Verstärkung interaktionsstarker Inhalte, Anonymität, Kontextkollaps (das Zusammenfallen ursprünglich getrennter sozialer Kontexte) sowie Echtzeitkommunikation die Sichtbarkeit normabweichender Beiträge begünstigen können (Schneiders, 2021, S. 286f.; Barth et al., 2023, S. 220ff.).

In der empirischen Hate-Speech-Forschung werden vor allem visuelle und interaktive Plattformen wie Instagram, Facebook, TikTok, YouTube und Twitter/X untersucht (Ruscher et al., 2024, S. 41f.; Schneiders, 2021, S. 286ff.). Diese Plattformen sind für die vorliegende Studie besonders relevant, weil Hate Speech und Counter Speech dort häufig in unmittelbarer Abfolge auftreten und soziale Normen situativ im Kommentarverlauf ausgehandelt werden.

Ferner weisen Studien auf dynamische Verstärkungseffekte hin. Hate Speech kann die Wahrscheinlichkeit weiterer abwertender Kommentare erhöhen und dadurch eskalierende Kommunikationsverläufe begünstigen (Madriaza et al., 2025, S. 38f.). Wiederholte Exposition kann außerdem dazu führen, dass Desensibilisierung und Veränderungen in der Wahrnehmung sozialer Normen gefördert werden. Dadurch wird feindselige Sprache als weniger außergewöhnlich oder problematisch empfunden (Soral et al., 2017, S. 9f.; Schäfer et al., 2023, S. 555ff.; Barth et al., 2023, S. 218f.). Dies unterstreicht die Relevanz, Plattformkontexte und Gruppenzugehörigkeit bei der Untersuchung von Hate Speech systematisch zu berücksichtigen.

Da Hate Speech marginalisierte Gruppen besonders betrifft, zugleich aber auch Wirkungen auf Mitlesende entfalten kann, ist eine getrennte Betrachtung der Teilnehmendengruppen (TG-M vs. TG-eM) zentral. Gleichzeitig erfordert das Instagram-Kommentarsetting eine realitätsnahe Abbildung, um valide Reaktionen auf Hate Speech und Counter Speech erfassen zu können (siehe Kapitel 5).

3.5 Wirkungen von Hate Speech

Empirische Forschung zeigt, dass Hate Speech sowohl unmittelbare psychische Belastungen als auch mittel- und langfristige soziale, normative und politische Veränderungen auslösen kann. Auf der Ebene direkt betroffener Personen ist Hate-Speech-Exposition mit erhöhtem psychischem Stress, Angst, depressiven Symptomen und Sicherheitsbedenken assoziiert (Khaleghipour et al., 2025, S. 2ff.; Saleem & Ramasubramanian, 2017, S. 374ff.). Studien berichten zudem über Rückzugs- und Vermeidungsverhalten in digitalen Räumen sowie erhöhte Wachsamkeit gegenüber weiterer Hate Speech (Gelber, 2019, S. 399ff., Khaleghipour et al., 2025, S. 1ff.). Befunde zeigen konsistent, dass identitätsbezogene Bedrohung mit psychischen Belastungsreaktionen und Vermeidungsverhalten einhergehen kann (Major & O'Brien, 2004, S. 411f.). Die Intensität dieser Belastungen steigt mit kumulativer Exposition und begrenzten individuellen Ressourcen (Bührer et al., 2024, S. 5; Gelber, 2019, S. 399ff.).

Neben Wirkungen auf direkt adressierte Personen zeigen Studien, dass Hate Speech auch Mitlesende beeinflussen kann. Die wiederholte Exposition gegenüber abwertenden

Kommentaren kann stereotypisierende Urteile und Vorurteile verstärken (Hsueh et al., 2015, S. 557ff.; Soral et al., 2017, S. 9f.).

Darüber hinaus beeinflusst Hate Speech die Wahrnehmung sozialer Normen in digitalen Öffentlichkeiten. Wiederholte Exposition kann sowohl deskriptive Normen (was als üblich wahrgenommen wird) als auch injunktive Normen (was als akzeptabel gilt) verschieben (Hsueh et al., 2015, S. 558ff.; Madriaza et al., 2025, S. 37ff., Bliuc et al., 2018, S. 81ff.; Castaño-Pulgarín et al., 2021, S. 1ff.). Solche Normverschiebungen können dazu beitragen, dass feindselige Sprache als weniger außergewöhnlich wahrgenommen wird und Hemmschwellen für weitere abwertende Kommentare sinken (Esau, 2023, S. 451f.).

Vor diesem Hintergrund stellt sich die Frage, welche Gegenmaßnahmen empirisch als wirksam diskutiert werden und unter welchen Bedingungen sie negative Effekte abschwächen. Der folgende Abschnitt bündelt hierzu zentrale Ansätze mit dem Schwerpunkt auf Counter Speech.

3.6 Gegenmaßnahmen mit Schwerpunkt Counter Speech

Hate Speech gilt in der Forschung als vielschichtiges Phänomen, dessen Eindämmung auf mehreren Ebenen erfolgt. Darunter regulatorische, technische, rechtliche, soziale und edukative Maßnahmen (Ruscher, 2024, 41ff.).

Rechtliche, regulatorische und technische Maßnahmen bilden einen bedeutenden institutionellen Rahmen, zielen allerdings primär auf die Begrenzung oder Entfernung problematischer Inhalte und weniger auf deren soziale Aushandlung. Staatliche Normen und plattformeigene Community-Standards erfüllen dabei vor allem eine symbolische Funktion, indem sie markieren, welche Ausdrucksformen als normwidrig gelten und gesellschaftlich nicht akzeptiert sind. Ihre Wirksamkeit hängt jedoch von konsequenter Durchsetzung ab und bleibt insbesondere bei subtilen oder kontextabhängigen Formen von Hate Speech begrenzt (Ruscher, 2024, S. 44f.).

Auch automatisierte Moderationssysteme stoßen an ihre Grenzen, da sie Ironie, implizite Bedeutungen oder gruppenspezifische Ausdrucksweisen nur unzureichend erfassen und dadurch sowohl Fehleinstufungen als auch Overblocking verursachen können (Ruscher, 2024, S. 42f.; Schäfer et al., 2023, S. 559). Automatisierte Counter-Speech-Bots zeigen bislang ebenfalls nur eingeschränkte Wirksamkeit, da ihnen kontextsensitive und glaubwürdige soziale Reaktionen fehlen (Ruscher, 2024, S. 43ff.; Hangartner et al., 2021).

Community- und Bildungsansätze legen hingegen mehr Wert auf Sensibilisierung, normative Orientierung und solidarische Eingriffe (Ruscher, 2024, S. 47ff.). Politische und zivilgesellschaftliche Initiativen machen deutlich, dass Hate Speech zunehmend als

gesellschafts- und demokratiepolitische Herausforderung verstanden werden muss und entsprechend kollektive Gegenstrategien diskutiert werden (ZARA, 2025; Europäische Kommission, 2023).

Eine zentrale Form solcher gemeinschaftlich getragenen Gegenstrategien ist Counter Speech. Sie umfasst alternative Narrative, die Hate Speech widersprechen, ohne primär auf die Bestrafung der sendenden Person abzuzielen (Ruscher, 2024, S. 43). Darüber hinaus kann Counter Speech Kommunikationsgrenzen markieren und so „norms of respectful interactions" stärken (Ruscher, 2024, S. 49). Zudem kann sie Betroffene unmittelbar unterstützen, indem soziale Zugehörigkeit signalisiert und negative emotionale Folgen abgemildert werden. Ruscher (2024) weist darauf hin, dass insbesondere bestätigende Reaktionen von Mitlesenden auf Counter Speech dazu beitragen können, „target's feelings of low safety and belongingness" zu verringern (S. 49).

Die empirische Evidenz zur Wirksamkeit von Counter Speech fällt insgesamt heterogen aus. Zwar zeigen mehrere Studien, dass Counter Speech unter bestimmten Bedingungen Normen stärken, Feindseligkeit relativieren oder Unterstützung für Betroffene signalisieren kann. Gleichzeitig finden andere Arbeiten keine Effekte oder berichten sogar gegenläufige Wirkungen, etwa Polarisierung oder verstärkte Distanz in politisch stark segmentierten Gruppen (Schäfer et al., 2023, S. 553ff.). Diese Uneindeutigkeit deutet darauf hin, dass Counter Speech keine universell wirksame Intervention ist. Vielmehr hängt ihre Wirkung stark davon ab, wie sie formuliert ist (z. B. aggressiv vs. unterstützend), wer sie äußert (CS-M vs. CS-eM) und in welchem Kontext sie rezipiert wird (z. B. Plattformlogiken, politische Voreinstellungen) (Hangartner et al., 2021, S. 1ff.; Ruscher, 2024, S. 49ff.).

Auch alternative Moderationsstrategien können je nach Kontext wirksamer sein. Álvarez-Benjumea und Winter (2018) zeigen etwa, dass moderates Löschen feindseliger Beiträge zukünftige Hate Speech stärker reduzieren kann als rein verbale Sanktionen (S. 10f.).

Die dargestellten Befunde verdeutlichen damit insgesamt eine starke Kontextabhängigkeit von Counter Speech. Im folgenden Abschnitt werden die zuvor theoretischen Einordnungen und Befunde kritisch zusammengeführt und daraus zentrale Forschungslücken abgeleitet.

3.7 Kritische Synthese und Forschungslücken

Die vorangegangenen Abschnitte zeigen, dass die empirische Forschung zu Hate Speech und Counter Speech in den letzten Jahren stark gewachsen ist, zugleich aber durch konzeptuelle und methodische Heterogenität geprägt bleibt (Barth et al., 2023, S. 213f.; Waqas et al., 2019,

S. 2; Bührer et al., 2024, S. 4f.). Die dargestellten Befunde verdeutlichen mehrere wiederkehrende Herausforderungen der bisherigen Forschung.

Erstens zeigt die Forschung zwar, dass Hate Speech unterschiedliche Reaktionen hervorrufen kann (Khaleghipour et al., 2025, S. 1ff.; Hsueh et al., 2015, S. 570ff.; Madriaza et al., 2025, S. 37ff.). Wirkmechanismen werden allerdings meist isoliert betrachtet oder in unterschiedlichen Designs untersucht. In der Folge bleibt häufig unklar, wie emotionale Reaktionen, kognitive Verarbeitungsprozesse und handlungsbezogene Bewertungen in konkreten Kommunikationssituationen zusammenwirken.

Zweitens bedarf es mehr experimenteller Forschung und einer stärkeren Berücksichtigung unterschiedlicher Kontexte. Reviews fordern „greater methodological diversity, particularly through experimental approaches and cross-country comparisons“ (Bührer et al., 2024, S. 8), sowie verstärkt Längsschnittdesigns und Vergleichsstudien. Der bisherige Schwerpunkt auf querschnittlichen Selbstauskünften limitiert insbesondere kausale Aussagen. Zudem erschwert die Konzentration auf WEIRD-Stichproben die Generalisierbarkeit (Castaño-Pulgarín et al., 2021, S. 2; Bührer et al., 2024, S. 9).

Drittens weist die Forschung zu Counter Speech ein ambivalentes Wirkprofil auf. Counter Speech kann Normsignale setzen, alternative Narrative anbieten und Betroffene unterstützen (Ruscher, 2024, S. 43ff.). Gleichzeitig zeigen Studien uneinheitliche, teils ausbleibende oder polarisierende Effekte, abhängig von Stil, Kontext oder individuellen Voreinstellungen (Schäfer et al., 2023, S. 553ff.; Álvarez-Benjumea & Winter, 2018, S. 10f.). Systematische Vergleiche zwischen Hate-Speech-Exposition mit und ohne Counter Speech in derselben Kommunikationssituation sowie kontrollierte Variationen der CS-Quelle liegen bislang nur begrenzt vor.

Viertens werden soziale Gruppenzugehörigkeiten und plattformspezifische Kommunikationslogiken häufig getrennt betrachtet oder nur unzureichend gemeinsam berücksichtigt. Obwohl Hate Speech marginalisierte Gruppen besonders betrifft und zugleich Wirkungen auf Mitlesende entfaltet, werden ethnische Minderheiten und die Mehrheitsgesellschaft selten vergleichend innerhalb eines Designs untersucht (Bührer et al., 2024, S. 5; Schäfer et al., 2023, S. 3). Ebenso bleiben plattformspezifische Kommunikationslogiken, etwa öffentliche Sichtbarkeit von Kommentaren oder unmittelbare Abfolgen von Hate Speech und Counter Speech, häufig unterbelichtet. Dennoch können sie Normwahrnehmungen und Interventionsentscheidungen beeinflussen (Barth et al., 2023, S. 220ff.).

Die vorliegende Studie erhebt nicht den Anspruch, sämtliche identifizierten Forschungslücken zu schließen, sondern fokussiert sich gezielt auf solche Aspekte, die im Rahmen eines experimentellen Designs unter kontrollierten Bedingungen empirisch zugänglich sind. Vor diesem Hintergrund zeigt sich eine zentrale Forschungslücke. Bislang liegen nur wenige experimentelle Untersuchungen vor, die Hate Speech und Counter Speech innerhalb realistischer Kommentarinteraktionen direkt vergleichen, mehrere Mechanismen gleichzeitig erfassen und dabei Gruppenzugehörigkeit sowie Plattformkontexte systematisch berücksichtigen. Auf Basis dieser Forschungslücken werden im folgenden Kapitel die Forschungsfrage und Hypothesen entwickelt. Diese sollen die Wirkungen von Hate Speech ohne Counter Speech versus Hate Speech mit Counter Speech sowie die Rolle der CS-Quelle empirisch prüfen.

4 Einführung in die vorliegende Studie

Aufbauend auf dem theoretischen Rahmen und der zusammengefassten empirischen Studienlage werden in diesem Kapitel die Hypothesen und die explorative Forschungsfrage der vorliegenden Studie hergeleitet. Der Fokus liegt auf den erwarteten Unterschieden zwischen den experimentellen Bedingungen sowie auf möglichen Moderationen durch CS-Quelle und Teilnehmendengruppen.

4.1 Zielsetzung und Forschungslogik der Studie

Die Studie ist so konzipiert, dass Wirkungen von Counter Speech unter kontrollierten Bedingungen kausal geprüft werden können. Durch die randomisierte Zuweisung zu einem realitätsnahen Instagram-Kommentarverlauf wird isoliert, ob und wie die Präsenz von Counter Speech die Wahrnehmung von Hate Speech verändert.

Zentral ist dabei eine zweistufige Vergleichslogik. Zunächst wird über einen geplanten Kontrast der Effekt der CS-Präsenz geprüft. Ein Kommentarverlauf mit Hate Speech ohne Counter Speech wird mit einem Verlauf verglichen, in dem auf Hate Speech sichtbar mit Counter Speech reagiert wird (HS vs. HS + CS). Damit wird untersucht, ob bereits die Existenz einer öffentlichen Gegenposition die Einordnung der Kommentare verändert.

Innerhalb der Bedingungen mit Counter Speech wird zusätzlich die CS-Quelle zwischen der Counter Speech von der Mehrheitsgesellschaft (CS-M) und der Counter Speech von der ethnischen Minderheit (CS-eM) verglichen. Dadurch kann geprüft werden, ob Counter Speech unterschiedlich wirkt, je nachdem, ob sie als gruppenübergreifendes Normsignal oder als Ausdruck betroffener Perspektive erscheint.

Ergänzend wird die Gruppenzugehörigkeit der Teilnehmenden berücksichtigt. Unterschieden wird zwischen der Teilnehmendengruppe der Mehrheitsgesellschaft (TG-M) und der Teilnehmendengruppe der ethnischen Minderheiten (TG-eM). Damit kann analysiert werden, ob Wirkungen auf Hate Speech und Counter Speech systematisch davon abhängen, welcher Teilnehmendengruppe Personen angehören.

Insgesamt erlaubt dieses Design die gleichzeitige Prüfung, ob Counter Speech die Reaktionen auf Hate Speech verändert und ob diese Effekte in Abhängigkeit von CS-Quelle und Teilnehmendengruppe variieren.

Im Folgenden werden Hypothesen und eine explorative Forschungsfrage entlang der drei abhängigen Variablen abgeleitet.

4.2 Wahrgenommene soziale Identitätsbedrohung

Im vorliegenden Kommentarsetting wird Hate Speech als öffentlich sichtbare gruppenbezogene Abwertung rezipiert. Ob diese Situation als bedrohlich für die eigene soziale Identität erlebt wird, dürfte davon abhängen, in welcher Beziehung Mitlesende zur adressierten Gruppe stehen. Für die Teilnehmendengruppe der ethnischen Minderheiten (TG-eM) ist eine höhere wahrgenommene soziale Identitätsbedrohung unter Hate Speech plausibel als für Teilnehmende der Mehrheitsgesellschaft (TG-M), da die Abwertung stärker auf die eigene soziale Zugehörigkeit beziehbar ist.

Die CS-Präsenz verändert den situativen Deutungsrahmen, weil die abwertende Äußerung nicht unwidersprochen im Kommentarverlauf stehen bleibt. Counter Speech macht sichtbar, dass die geäußerte Abwertung nicht sozial geteilt wird. Entsprechend ist zu erwarten, dass die wahrgenommene soziale Identitätsbedrohung bei Hate Speech mit Counter Speech geringer ist als bei Hate Speech ohne Counter Speech. Dieser Reduktionseffekt sollte insbesondere bei TG-eM ausgeprägt sein, da hier die Ausgangsbedrohung höher anzunehmen ist.

Ferner ist plausibel, dass die Wirkung von Counter Speech von der CS-Quelle abhängt. Counter Speech aus der Mehrheitsgesellschaft (CS-M) kann als gruppenübergreifendes Norm- und Solidaritätssignal interpretiert werden, während Counter Speech aus der ethnischen Minderheit (CS-eM) als Ausdruck betroffener Erfahrung und Authentizität wahrgenommen werden kann. Welche Deutung in den Vordergrund tritt, dürfte kontextabhängig sein. Daraus ergeben sich folgende Hypothesen:

H1 (wahrgenommene soziale Identitätsbedrohung): Die wahrgenommene soziale Identitätsbedrohung ist unter Hate Speech ohne Counter Speech höher als nach Hate Speech mit Counter Speech. Dieser Effekt der CS-Präsenz fällt bei der Teilnehmendengruppe ethnischer Minderheiten (TG-eM) stärker aus als bei der Teilnehmendengruppe der Mehrheitsgesellschaft (TG-M).

H1a (CS-Quelle): Unter den Bedingungen mit Counter Speech unterscheidet sich die wahrgenommene soziale Identitätsbedrohung in Abhängigkeit von der CS-Quelle (CS-M vs. CS-eM). Dieser Unterschied wird durch die Teilnehmendengruppen (TG-M vs. TG-eM) moderiert.

4.3 Stereotypbezogene Zuschreibungen (Eigenschaftszuschreibung und Distanz/Vermeidung)

Wie in Kapitel 2.4.2 dargestellt, kann Hate Speech kurzfristig stereotypbezogene Wissens- und Bewertungsschemata aktivieren und dadurch die Wahrscheinlichkeit erhöhen, dass

Mitlesende in anschließenden Bewertungen auf negative Gruppenbilder zurückgreifen. Im Rahmen der vorliegenden Studie werden stereotypbezogene Zuschreibungen als kognitive Wirkungsebene erfasst. Im Fokus steht dabei nicht die langfristige Entstehung neuer Stereotype, sondern die kurzfristige Aktivierung und Anwendung wertender gruppenbezogener Zuschreibungen im Anschluss an die Exposition.

Die CS-Präsenz kann diesem Prozess theoretisch entgegenwirken, indem sie alternative Deutungsangebote bereitstellt, Verzerrungen zurückweist oder korrigiert und zugleich normative Grenzen markiert. Dadurch könnte die Anwendbarkeit aktivierter negativer Gruppenurteile reduziert werden. Für die Teilnehmendengruppe Mehrheitsgesellschaft (TG-M) ist dies besonders relevant, da die in Hate Speech adressierte Gruppe typischerweise eine Fremdgruppe darstellt und negative Zuschreibungen hier eher als Bewertungsgrundlage verfügbar sein können.

Auch bei dieser abhängigen Variable ist ein Einfluss der CS-Quelle plausibel. Counter Speech aus der ethnischen Minderheit (CS-eM) kann als erfahrungsnahe Perspektive angesehen werden, während Counter Speech aus der Mehrheitsgesellschaft (CS-M) eher als ein allgemeines Normsignal betrachtet werden kann. Welche dieser Deutungen die Anwendung stereotypbezogener Zuschreibungen stärker beeinflusst, dürfte zwischen TG-M und TG-eM variieren. Auf dieser Grundlage werden folgende Hypothesen formuliert:

H2 (Stereotypbezogene Zuschreibungen): Die stereotypbezogene Zuschreibung ist unter Hate Speech ohne begleitende Counter Speech höher als nach Hate Speech mit Counter Speech. Dieser Effekt der CS-Präsenz fällt bei der Teilnehmendengruppe der Mehrheitsgesellschaft (TG-M) stärker aus, als bei der Teilnehmendengruppe ethnischer Minderheit (TG-eM).

H2a (CS-Quelle): Unter den Bedingungen mit Counter Speech unterscheiden sich stereotypbezogene Zuschreibungen in Abhängigkeit von der CS-Quelle (CS-M vs. CS-eM). Dieser Unterschied wird durch die Teilnehmendengruppen (TG-M vs. TG-eM) moderiert.

4.4 Interventionsbereitschaft

Während wahrgenommene soziale Identitätsbedrohung und stereotypbezogene Zuschreibungen beschreiben, wie Hate Speech wahrgenommen und verarbeitet wird, richtet sich der Fokus dieser Wirkungsebene auf handlungsbezogene Reaktionen im Kommentarsetting. Erfasst wird die Bereitschaft Teilnehmenden, selbst aktiv auf Hate Speech zu reagieren. Interventionsbereitschaft wird im Sinne des Bystander-Modells über Schweregradwahrnehmung, Verantwortungsgefühl und Interventionsabsicht operationalisiert.

Die CS-Präsenz kann diese Bereitschaft in unterschiedlicher Weise beeinflussen. Einerseits kann Counter Speech als sichtbares Signal dienen, dass die geäußerte Abwertung nicht unwidersprochen bleibt, und damit die Wahrnehmung sozialer Normen verändern sowie eigene Interventionsabsichten fördern. Andererseits ist es möglich, dass bereits vorhandene Counter Speech im Kommentarverlauf den Eindruck erzeugt, eine Reaktion sei nicht mehr notwendig, wodurch die eigene Interventionsbereitschaft sinken könnte.

Die empirische Forschung zeigt hierzu bislang kein einheitliches Muster, sondern kontextabhängige Befunde (siehe Kapitel 3.6). Vor diesem Hintergrund ergeben sich für die vorliegende Studie keine eindeutig gerichteten Annahmen zur Wirkung der CS-Präsenz auf die Interventionsbereitschaft. Stattdessen wird mit einer explorativen Forschungsfrage offen geprüft, ob und in welche Richtung sich Unterschiede zwischen Hate Speech ohne Counter Speech und Hate Speech mit Counter Speech zeigen. Zudem wird untersucht, ob sich diese möglichen Effekte zwischen den Teilnehmendengruppen (TG-M vs. TG-eM) unterscheiden. Daraus wird folgende explorative Forschungsfrage abgeleitet:

> **EF1 (Interventionsbereitschaft):** Wie beeinflusst die Exposition von Hate Speech ohne Counter Speech im Vergleich zu Hate Speech mit Counter Speech die Interventionsbereitschaft der Teilnehmenden, und unterscheiden sich diese Effekte zwischen den Teilnehmendengruppen (TG-M vs. TG-eM)?

4.5 Soziodemografische- und Moderatorvariablen

Zur Beschreibung der Stichprobe und zur Prüfung möglicher Unterschiede zwischen den experimentellen Bedingungen werden soziodemografische Variablen erfasst. Dazu zählen Alter, Geschlecht, Bildungsniveau und Wohnort. Diese Variablen dienen der Einordnung der Stichprobe sowie der Kontrolle potenzieller Verzerrungen zwischen den Experimentalbedingungen. Die Gruppenzugehörigkeit der Teilnehmenden (TG-M vs. TG-eM) stellt dabei keinen Kontrollfaktor dar, sondern ist ein zentraler untersuchter Personenfaktor im Design der Studie.

Ergänzend werden explorative Moderatorvariablen berücksichtigt, um theoriegeleitet zu prüfen, ob und unter welchen Bedingungen sich die Wirkung von Hate Speech und Counter Speech unterscheiden. Damit wird an Befunde angeschlossen, die auf kontext- und personenbezogene Unterschiede in der Verarbeitung solcher Inhalte hinweisen (siehe Kapitel 3). Erfasst werden die politische Orientierung, vorbestehende Einstellungen gegenüber ethnischen Minderheiten und die Social-Media-Nutzungsintensität sowie der Social-Media-Nutzungstyp (Informationssuche, Beiträge kommentieren, Informationen teilen). Die

Einbeziehung dieser Variablen hilft dabei, unterschiedliche Wirkungsmuster zu erkennen und die Ergebnisse im Hinblick auf bestehende Forschungslücken genauer einzuordnen.

4.6 Konzeptionelles Modell und Hypothesenstruktur

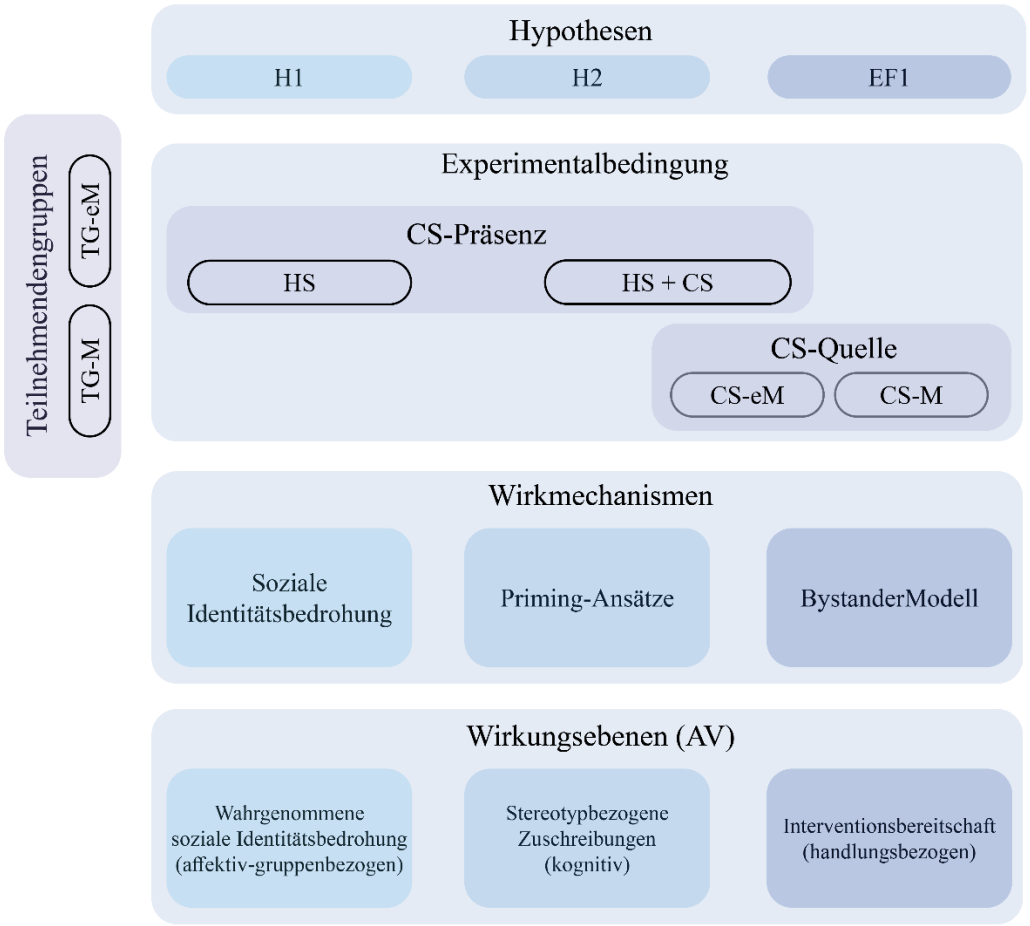
Die Hypothesen der vorliegenden Studie beziehen sich auf die Rolle der CS-Präsenz, CS-Quelle und Teilnehmendengruppen (TG-M vs. TG-eM). Diese Faktoren werden in Bezug auf drei Wirkungsebenen untersucht: wahrgenommene soziale Identitätsbedrohung, stereotypbezogene Zuschreibungen und Interventionsbereitschaft.

Für die wahrgenommene soziale Identitätsbedrohung und die stereotypbezogenen Zuschreibungen werden gerichtete Annahmen formuliert. Für die Interventionsbereitschaft wird aufgrund möglicher gegenläufiger Effekte eine explorative Forschungsfrage untersucht.

Abbildung 2 veranschaulicht die Zusammenhänge zwischen den experimentellen Faktoren, den berücksichtigten Teilnehmendengruppen sowie den drei untersuchten Wirkungsebenen und dient als konzeptionelle Orientierung für die folgenden Kapitel.

Abbildung 2

Erweitertes konzeptionelles Modell zu Wirkungsebenen von Hate Speech und Counter Speech



5 Methodik

Dieses Kapitel beschreibt das methodische Vorgehen der vorliegenden Studie. Aufbauend auf dem theoretischen Rahmen (Kapitel 2), dem empirischen Forschungsstand (Kapitel 3) sowie den daraus abgeleiteten Hypothesen und der explorativen Forschungsfrage (Kapitel 4) werden nun Studiendesign, Stichprobe und Stimulusmaterial dargestellt. Zudem werden das Erhebungsinstrument, die Operationalisierung der Variablen sowie forschungsethische Aspekte präsentiert. Ziel ist eine transparente und nachvollziehbare Dokumentation des Vorgehens.

5.1 Studiendesign

Zur Beantwortung der Forschungsfrage und zur Überprüfung der Hypothesen sowie der explorativen Forschungsfrage wurde ein quantitatives Online-Experiment im Zwischen-Gruppen-Design durchgeführt. Die randomisierte Zuweisung der Teilnehmenden zu den Experimentalbedingungen sowie die standardisierte Stimuluspräsentation ermöglichten Schlussfolgerungen über die Effekte. Die Teilnehmendengruppe (TG-M vs. TG-eM) wurde nicht randomisiert, sondern per Selbstidentifikation erhoben und als Personenfaktor berücksichtigt.

Die Experimentalbedingung wurde in drei Ausprägungen variiert. Hate Speech ohne Counter Speech (HS), Hate Speech mit Counter Speech aus der Mehrheitsgesellschaft (HS + CS-M) sowie Hate Speech mit Counter Speech aus der ethnischen Minderheit (HS + CS-eM). Für Analysen zur CS-Präsenz wurden die beiden Counter-Speech-Bedingungen zu einer Kategorie zusammengefasst (HS vs. HS + CS). Effekte der CS-Quelle wurden innerhalb der Teilstichprobe mit Counter Speech durch den direkten Vergleich von HS + CS-M und HS + CS-eM geprüft. Die Teilnehmendengruppe wurde über Selbstidentifikation erfasst und für Analysen als TG-M vs. TG-eM berücksichtigt.

5.2 Stichprobe und Rekrutierung

Die Rekrutierung der Teilnehmenden erfolgte im Zeitraum vom 09.06.2025 bis 12.07.2025 über soziale Medien (u. a. Instagram, WhatsApp) sowie über persönliche und berufliche Kontakte. Einschlusskriterien waren ein Mindestalter von 18 Jahren sowie ein Wohnsitz im deutschsprachigen Raum (Deutschland, Österreich oder Schweiz). Ein finanzieller oder materieller Anreiz wurde nicht bereitgestellt. Ziel war eine heterogene Stichprobe mit einer geplanten Größe von $N \approx 200$, um neben den Haupteffekten auch Interaktionseffekte statistisch prüfen zu können.

5.3 Stimulusmaterial und Plattformkontext

Das Stimulusmaterial bestand aus einem Instagram-Post mit zugehörigem Kommentarbereich. Der Post sowie der Hate-Speech-Kommentar waren in allen Experimentalbedingungen identisch. Variiert wurde ausschließlich die Counter Speech. Dabei wurde zum einen die CS-Präsenz umgesetzt (Hate Speech ohne Counter Speech vs. Hate Speech mit Counter Speech). Zum anderen wurde innerhalb der Bedingungen mit Counter Speech die CS-Quelle variiert (Counter Speech aus der Mehrheitsgesellschaft vs. aus der ethnischen Minderheit). Dadurch konnten Unterschiede in den abhängigen Variablen auf diese experimentellen Variationen zurückgeführt werden.

Der Hate-Speech-Kommentar wurde in Anlehnung an die Arbeitsdefinition in Kapitel 2.1 gestaltet und zur inhaltlichen Konkretisierung entlang der empirisch etablierten Inhaltsdimensionen operationalisiert (Kapitel 3.3). Konkret wurden drei in der Forschung häufig identifizierte Dimensionen umgesetzt. Der Kommentar enthielt Elemente negativer Stereotypisierung („faule Schmarotzer“), Dehumanisierung („dreckiges Pack“) sowie Ausgrenzungs- bzw. Vernichtungsbezug („Schmeißt sie raus“). Damit konnte Gruppenbezug, Abwertung bzw. Gewalt- und Ausgrenzungslegitimation sowie Schädigungspotenzial im Stimulus realitätsnah abgebildet werden.

Die Counter Speech wurde in Übereinstimmung mit der Arbeitsdefinition in Kapitel 2.3 als öffentlich sichtbare Antwort auf Hate Speech gestaltet, die die Hassäußerung explizit zurückweist („Rassismus geht gar nicht!“) und eine solidarische bzw. unterstützende Position einnimmt („Jeder verdient Respekt, Punkt.“), ergänzt um Hashtags und Emojis als typische Plattformmarker. Die beiden Counter-Speech-Varianten unterschieden sich ausschließlich in der CS-Quelle, die über eine explizite Selbstpositionierung im Kommentartext hergestellt wurde. In der CS-Quelle (ethnische Minderheit) wurde Betroffenheit und Erfahrung betont („Ich erlebe Rassismus“), in der CS-Quelle (Mehrheitsgesellschaft) eine gruppenübergreifende Normsetzung („Ich gehöre zur Mehrheit“). Tonalität, Argumentationsstruktur und Länge wurden konstant gehalten, um die Effekte der CS-Quelle isoliert untersuchen zu können.

Die Stimuli wurden in ein Instagramähnliches Layout eingebettet (Post und Kommentarbereich). Profilbilder und Nutzer*innennamen wurden aus Neutralitätsgründen unkenntlich gemacht. Vor der Haupterhebung wurde ein Pretest zur Verständlichkeit durchgeführt (n = 7), um Verständlichkeit, Glaubwürdigkeit, Belastungsangemessenheit sowie die technische Darstellung der Stimuli zu prüfen. Auf Grundlage der Rückmeldungen wurden kleinere sprachliche Anpassungen und Layoutkorrekturen vorgenommen. Zudem

wurde das Debriefing präzisiert, indem die dargestellten Inhalte explizit als fiktiv und ausschließlich zu Forschungszwecken erstellt gekennzeichnet wurden. Die drei Experimentalbedingungen mit den konkreten Formulierungen von HS, CS-eM und CS-M sind in den Abbildungen 3, 4 und 5 dargestellt.

Abbildung 3

HS

Abbildung 4

HS + CS-eM

Abbildung 5

HS + CS-M

5.4 Erhebungsinstrument und Ablauf

Die Datenerhebung erfolgte mittels eines standardisierten Online-Fragebogens über SoSciSurvey. Das Onlineformat ermöglichte eine standardisierte Präsentation der Stimuli und randomisierte Zuweisung zu den Bedingungen. Zudem entsprach die Online-Rezeption der Stimuli dem typischen Nutzungskontext von Instagram-Kommentarbereichen und ermöglichte eine anonyme Teilnahme, wodurch Verzerrungen durch soziale Erwünschtheit bei sensiblen Themen tendenziell reduziert wurden.

Der Fragebogen war in sechs Abschnitte gegliedert. Zu Beginn wurde ein Briefing (Abbildung C1 im Anhang) sowie die Einwilligungserklärung mit Hinweis auf potenziell belastende Inhalte präsentiert (Abbildung C2 im Anhang). Anschließend wurden zentrale Moderatorvariablen erhoben, darunter Einstellungen gegenüber ethnischen Minderheiten,

politische Orientierung und Social-Media-Nutzung (Abbildung C3 im Anhang). Darauf folgte die Stimuluspräsentation (vgl. Abbildung 3, 4 und 5). Zur Sicherung der Aufmerksamkeitsvalidität wurde nach der Stimuluspräsentation eine Kontrollfrage zum Stimulusinhalt integriert (Abbildung C4 im Anhang). Im direkten Anschluss wurden die abhängigen Variablen (wahrgenommene soziale Identitätsbedrohung, stereotypbezogene Zuschreibungen, Interventionsbereitschaft) erfasst (Abbildung C5 im Anhang). Es folgten soziodemografische Angaben zu Alter, Geschlecht, Bildungsabschluss, Wohnort und ethnischer Selbstidentifikation (Abbildung C6 im Anhang). Den Abschluss bildete ein Debriefing mit Hinweisen auf Unterstützungs- und Kontaktstelle (Abbildung C7 im Anhang). In einem eingeblendeten Erläuterungsfeld wurde folgende Definition bereitgestellt, da der Begriff „ethnische Minderheiten“ für manche Teilnehmende potenziell abstrakt oder unterschiedlich interpretierbar sein kann und so ein gemeinsames Verständnis unterstützt werden sollte: „Unter ethnischen Minderheiten verstehen sich Bevölkerungsgruppen, die sich durch gemeinsame ethnische Merkmale wie Herkunft, Sprache, Kultur oder Religion von der Mehrheitsgesellschaft unterscheiden und in der Gesellschaft in der Regel eine zahlenmäßig kleinere Gruppe darstellen. Dazu zählen häufig Menschen mit Migrationshintergrund oder ausländischer Staatsangehörigkeit, sofern sie sich als Teil einer ethnischen Minderheit verstehen oder gesellschaftlich so wahrgenommen werden.“.

5.5 Operationalisierung der Variablen

Erhoben wurden drei abhängige Variablen: wahrgenommene soziale Identitätsbedrohung, stereotypbezogene Zuschreibungen (Eigenschaftszuschreibung, Distanz/Vermeidung) sowie Interventionsbereitschaft (Schweregradwahrnehmung, Verantwortungsgefühl, Interventionsabsicht). Die Skalen im Fragebogen stammten aus forschungsnahen Studien, wurden ins Deutsche übersetzt und an den Stimuluskontext angepasst.

5.5.1 Abhängige Variablen

Wahrgenommene soziale Identitätsbedrohung (D101)

Die wahrgenommene soziale Identitätsbedrohung bildete die affektiv-gruppenbezogene Wirkungsebene der Studie ab. Die Messung erfolgte in Anlehnung an Van Houtven et al. (2024), die die *perceived identity threat* mit fünf Items erfassen (S. 932). Sie beziehen diese Items auf Skalenbestandteile von Leonhard et al. (2018) sowie Breakwell und Jaspal (2021). Die fünf Items wurden ins Deutsche übersetzt und auf das Kommentarsetting angepasst („Nach dem Lesen eines oder mehrerer Kommentare unter dem Instagram-Post fühle ich mich... erniedrigt, diskriminiert, beleidigt, bedroht, in meinem Selbstwertgefühl untergraben”). Die Antworten wurden auf einer 5-Punkte-Likert-Skala erhoben (1 = trifft

überhaupt nicht zu bis 5 = trifft voll zu). Höhere Werte zeigen somit eine stärkere wahrgenommene soziale Identitätsbedrohung an.

Stereotypbezogene Zuschreibungen (D102, D103)

Die kognitive Wirkungsebene wurde in der Studie als stereotypbezogene Eigenschaftszuschreibung und Distanz/Vermeidung erfasst. Für die vorliegende Studie wurden beide Skalen ins Deutsche übertragen und inhaltlich auf die im Stimulus adressierte Gruppe (ethnische Minderheiten) bezogen. Eigenschaftszuschreibung wurde über vier Items operationalisiert (offen, tolerant, freundlich, vertrauenswürdig) und theoretisch in der Forschung zu Intergruppenwahrnehmung verankert (Deaux et al., 2007; Stephan et al., 1998; Van Houtven et al., 2024, S. 932). Ergänzend wurde Distanz als Vermeidungsdimension über drei Items erfasst („Ich möchte nichts mit … zu tun haben“, „Ich meide...“, „Ich halte Abstand...“), die Van Houtven et al. (2024) aus Mackie et al. (2000) adaptieren (S. 932).

Interventionsbereitschaft (D104-D106)

Die Interventionsbereitschaft bildet die handlungsbezogene Wirkungsebene der Studie ab (siehe Kapitel 2.4.3) und wurde in Anlehnung an Obermaier et al. (2014) operationalisiert, die Prozesse des Bystander-Modells in einem experimentellen Social-Media-Setting untersuchten. Obermaier et al. (2014) erfassten Interventionsbereitschaft mehrdimensional über die Schweregradwahrnehmung (*degree of severity*), ein Gefühl persönlicher Verantwortung zur Intervention (*feeling of responsibility*, übernommen aus Greitemeyer et al. (2006) adaptiert), sowie die Interventionsabsicht (*intention to intervene*) (S. 1495). Für die vorliegende Studie wurden diese drei Messdimensionen ins Deutsche übertragen und auf das Instagram-Kommentarsetting angepasst. Tabelle 1 fasst die abhängigen Variablen, ihre theoretische Verortung und die eingesetzten Messinstrumente zusammen.

Tabelle 1

Übersicht der abhängigen Variablen und ihrer Operationalisierung

H, EF	Konstrukt	Abhängige Variablen	Ursprüngliche Skala	Angepasste Skala
H1, H1a	Soziale Identitätsbedrohung (affektiv-gruppenbezogen)	Wahrgenommene soziale Identitätsbedrohung	(1) After reading the comment I feel degraded. (2) After reading the comment I feel discriminated. (3) After reading the comment I feel offended. (4) After reading the comment I feel threatened. (5) After reading the comment I feel that my sense of self-worth is undermined. 5-Punkte-Likert-Skala (Van Houtven et al., 2024, S. 947, 932)	D101: Nach dem Lesen eines oder mehrerer Kommentare unter dem Instagram-Post fühle ich mich... erniedrigt, diskriminiert, beleidigt, bedroht, in meinem Selbstwertgefühl untergraben. Kreuzen Sie jeweils die Antwort an, die am besten auf Ihr Empfinden zutrifft. 5-Punkte-Likert-Skala: 1 = trifft überhaupt nicht zu bis 5 = trifft voll zu
H2, H2a	Priming (kognitiv)	Stereotypbezogene	Out-Group Attitude:	D102: Wenn Sie an ethnische Minderheiten denken, inwieweit glauben Sie, dass die

		Zuschreibungen (Eigenschaftszuschreibung)	When you think of (heterosexual/White) people in general, to what extent do you believe that the following characteristics apply to them? In general, (heterosexual/ White) people are: (1) Open (2) Tolerant (3) Friendly (4) Trustworthy 5-Punkte-Likert-Skala (Van Houtven et al., 2024, S. 947)	folgenden Eigenschaften auf diese Gruppe zutreffen? (Offen, tolerant, freundlich, vertrauenswürdig) Bitte geben Sie Ihre persönliche Einschätzung an. 5-Punkte-Likert-Skala: 1 = trifft überhaupt nicht zu bis 5 = trifft voll zu
H2, H2a	Priming (kognitiv)	Stereotypbezogene Zuschreibung (Distanz/Vermeidung)	Out-Group Avoidance: (1) I do not want to have anything to do with (heterosexual/White) individuals. (2) I generally avoid (heterosexual/White) people. (3) I generally keep my distance with (heterosexual/White) individuals. 5-Punkte-Likert-Skala (Van Houtven et al., 2024, S. 947)	D103: Inwieweit stimmen Sie den folgenden Aussagen zu? Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen.“ „(Ich möchte nichts mit ethnischen Minderheiten zu tun haben. Ich meide ethnische Minderheiten, ich halte Abstand zu ethnischen Minderheiten) 5-Punkte-Likert-Skala: 1 = Ich stimme überhaupt nicht zu bis 5 = Ich stimme voll zu
EF1	Bystander-Modell (handlungsbezogen)	Interventionsbereitschaft (Schweregradwahrnehmung)	Degree of severity of the cyberbullying incident (not threatening - very threatening, not alarming - very alarming, not dangerous - very dangerous) 5-Punkte-Likert-Skala (Obermaier et al., 2014)	D104: Bitte bewerten Sie, inwieweit Sie den unter dem Instagram-Post gezeigten Kommentar als bedrohlich, alarmierend oder gefährlich empfinden. Nicht bedrohlich - Sehr bedrohlich. Nicht alarmierend - Sehr alarmierend. Nicht gefährlich - Sehr gefährlich. 5-Punkte-Likert-Skala
EF1	Bystander-Modell (handlungsbezogen)	Interventionsbereitschaft (Gefühl der persönlichen Verantwortung)	The feeling of responsibility to intervene in the cyberbullying incident (I highly feel personally responsible to support Michi, I consider it as my duty to help Michi, and It is my obligation to do something about this comment) 5-Punkte-Likert-Skala (Obermaier et al., 2014)	D105: Inwieweit stimmen Sie den folgenden Aussagen zu? Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen. Ich fühle mich persönlich stark verantwortlich, ethnische Minderheiten zu unterstützen; Ich betrachte es als meine Pflicht, ethnischen Minderheiten zu helfen; Es ist meine Verpflichtung etwas gegen den gezeigten Kommentar zu unternehmen. 5-Punkte-Likert-Skala: 1 = Ich stimme überhaupt nicht zu bis 5 = Ich stimme voll zu
EF1	Bystander-Modell (handlungsbezogen)	Interventionsbereitschaft (Interventionsabsicht)	Intention to intervene in the cyberbullying incident (I would intervene) 5-Punkte-Likert-Skala (Obermaier et al., 2014)	D106: Bitte geben Sie an, inwieweit Sie der folgenden Aussage zustimmen. Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen. Ich würde auf den Kommentar reagieren. 5-Punkte-Likert-Skala: 1 = Ich stimme überhaupt nicht zu bis 5 = Ich stimme voll zu

5.5.2 Soziodemografische- und Moderatorvariablen

Zur Berücksichtigung potenzieller Moderator- und Einflussfaktoren wurden Einstellungen gegenüber ethnischen Minderheiten, politische Orientierung sowie Social-Media-Nutzung erfasst (siehe Kapitel 4.4). Moderatorvariablen wurden vor der Stimuluspräsentation erhoben, um sicherzustellen, dass sie nicht durch die experimentelle Exposition verzerrt wurden.

Einstellungen gegenüber ethnischen Minderheiten (B101). Zur Erfassung vorbestehender Einstellungen gegenüber ethnischen Minderheiten wurde ein Single-Item

eingesetzt, angelehnt an Schäfer et al. (2023). In der Originalstudie erfassten Schäfer et al. (2023) *pre-existing attitudes* zu mehreren sozialen Gruppen, indem Teilnehmende ihre allgemeine Wahrnehmung der Gruppen auf einer Skala von −5 (very negative) bis +5 (very positive) angaben. Die Autor*innen orientierten sich dabei an etablierten Vorarbeiten, in denen gruppenbezogene Einstellungen ebenfalls über Single-Item-Messungen erhoben wurden (u.a. Küpper et al., 2017) (Schäfer et al., 2023, S. 563). In der vorliegenden Studie wurde das Item ins Deutsche übertragen und auf den Untersuchungsgegenstand angepasst.

Politische Orientierung (B102). Die politische Orientierung wurde über eine Selbstverortung auf einer Links-Rechts-Skala erfasst (0 = links bis 100 = rechts), ebenfalls angelehnt an Schäfer et al. (2023). Die Erhebung der politischen Ausrichtung kann als zentrale Kontrollvariable begründet werden, da politische Selbstpositionierung eng mit Einstellungen zu sozialen Gruppen sowie mit der Bewertung gesellschaftlicher Konfliktthemen zusammenhängt (Chambers et al., 2013; Schäfer et al., 2023, S. 567). Dies kann sowohl die Bewertung von Hate Speech als auch die Akzeptanz und Wirksamkeit von Counter Speech beeinflussen.

Social-Media-Nutzungsintensität (B105). Da die alltägliche Social-Media-Exposition die Gewöhnung an Kommentaröffentlichkeiten sowie die Wahrscheinlichkeit einschlägiger Kontakterfahrungen mit Hate Speech beeinflussen kann, wurde zusätzlich die durchschnittliche tägliche Nutzungsdauer erhoben. Die Kategorisierung (unter 1 Stunde, 1–2 Stunden, 2–3 Stunden, 3–4 Stunden, 4 Stunden und mehr) orientierte sich an Wang (2018, S. 26), der Social-Media-Nutzungszeit in vergleichbaren Stundenkategorien erfasste.

Social-Media-Nutzungstyp (B103). In Anlehnung an Jia & Schumann (2023) wurde erfasst, wie häufig Teilnehmende soziale Medien nutzen, um Informationen zu suchen, Beiträge zu kommentieren und Inhalte zu teilen. Das Antwortformat entsprach der Originalskala (1 = nie bis 5 = sehr oft) (S. 13). Dadurch wurden Unterschiede zwischen eher passiver und aktiver Social-Media-Nutzung berücksichtigt, die als kontextrelevante Einflussfaktoren für die Verarbeitung von Kommentarinhalten galten (Jia & Schumann, 2023, S. 13).

Kontrollfrage (B201). Zur Sicherung der Datenqualität wurde eine Aufmerksamkeitskontrollfrage integriert, die den Inhalt des Instagram-Posts abfragt.

Tabelle 2 zeigt eine Übersicht der Moderator-, Kontroll- und Kontextvariablen sowie deren Operationalisierung in der vorliegenden Studie.

Tabelle 2

Übersicht der Moderator- und Kontrollvariablen und ihre Operationalisierung

Variable	Ursprüngliche Skala	Angepasste Skala
Voreinstellung gegenüber ethnischen Minderheiten	Pre-ExistingAttitudes: If you think about the following groups, is your general attitude about them rather negative or positive? To answer this question, you can use a scale from -5 very negative to +5 very positive. With the values in between, you can graduate your answer. -5 (very negative), 0 (neutral), +5 (very positive) 5-Punkte-Likert-Skala (Schäfer et al., 2023, S. 563)	B101 Wenn Sie an ethnische Minderheiten im deutschsprachigen Raum denken, wie würden Sie Ihre allgemeine Einstellung gegenüber dieser Gruppe beschreiben? Bitte geben Sie an, wie positiv oder negativ Sie gegenüber ethnischen Minderheiten eingestellt sind. Verwenden Sie dazu eine Skala von -5 (sehr negativ) bis +5 (sehr positiv). Mit den Werten dazwischen können Sie Ihre Meinung entsprechend abstufen. Skala: -5 = sehr negativ bis +5 = sehr positiv
Politische Voreinstellung	In politics, it is often about being „left“ or „right“. Where would you see yourself on a scale on which 0 is left and 10 is right? If you don't want to answer this question, please proceed with the next one. (Schäfer et al., 2023, S. 567)	B102 In der Politik geht es oft darum, links oder rechts zu stehen. Wo würden Sie sich auf einer Skala von 0 (links) bis 10 (rechts) einordnen?
Social-Media-Nutzungsintensität	Average time social media was used daily (in hours): less than one hour, 1 to 2 hours, 2 to 3 hours, 3 to 4 hours more than 4 hours (Wang, 2018, S. 26)	B105 Wie viel Zeit verbringen Sie durchschnittlich pro Tag auf sozialen Medien? Antwortmöglichkeiten: Unter 1 Stunde, 1 bis 2 Stunden, 2 bis 3 Stunden, 3 bis 4 Stunden, 4 Stunden und mehr
Social-Media-Nutzungstyp	When using social media, how often do you ..? Browse information; Comment on posts; share information. (Never, rarely, sometimes, often, very often) (Jia & Schumann, 2023, S. 13)	Wie oft machen Sie Folgendes, wenn Sie soziale Medien nutzen? Wählen Sie jeweils die Antwort aus, die Ihrem Verhalten am besten entspricht. (Informationen suchen, Beiträge kommentieren, Informationen teilen) (nie, selten, manchmal, oft, sehr oft)
Aufmerksamkeitsfrage	–	In dem Instagram-Post ging es um. (Ethnische Minderheiten Wahlverhalten in Deutschland)

Soziodemografische Variablen. Im Anschluss an die Erhebung der zentralen Variablen wurden soziodemografische Angaben erfasst, um die Stichprobe zu beschreiben und potenzielle Einflüsse berücksichtigen zu können. Erhoben wurden Alter (freie Eingabe in Jahren), Geschlecht (männlich/weiblich/divers/keine Angabe), Bildungsabschluss (kategorial) sowie Wohnort (Deutschland/Österreich/Schweiz). Die Gruppenzugehörigkeit wurde über ethnische Selbstidentifikation bestimmt („Fühlen Sie sich einer ethnischen Minderheit im deutschsprachigen Raum zugehörig?“, mit Antwortoptionen Ja/Nein/Keine Angabe).

Die Erfassung über Selbstidentifikation ist methodisch und ethisch begründet, da soziale Identität nicht extern zugeschrieben werden sollte, sondern durch die Befragten selbst definiert wird. Wie Sharghi et al. (2024) betonen, „the participant is the world's expert on their identity“ (S. 4).

Ergänzend wurde ein optionales Freitextfeld angeboten, um den Teilnehmenden Raum für zusätzliche Anmerkungen zu geben.

5.6 Analytisches Vorgehen und Auswertungsplan

Die Datenaufbereitung und statistische Auswertung erfolgten mit IBM SPSS Statistics. Sämtliche Analyseschritte wurden syntaxbasiert dokumentiert, um Transparenz und Reproduzierbarkeit zu gewährleisten. Das Signifikanzniveau wurde auf $\alpha = .05$ festgelegt. Zusätzlich wurden Effektstärken berichtet (partielles η^2 bei Varianzanalysen, Cohen's d bei Mittelwertvergleichen).

Vor den Hypothesentests wurde der Datensatz anhand vorab definierter Kriterien bereinigt. Ausgeschlossen wurden abgebrochene Teilnahmen, fehlende Einwilligung bzw. fehlende Bestätigung der Volljährigkeit sowie Fälle mit falsch beantworteter Aufmerksamkeitskontrolle zum Stimulus. Fälle ohne Angabe zur ethnischen Selbstidentifikation wurden in Analysen, die die Teilnehmendengruppen berücksichtigen, nicht einbezogen.

Für alle mehrdimensionalen Skalen wurden Skalenmittelwerte berechnet. Die interne Konsistenz der Skalen wurde anhand von Cronbach's α geprüft.

Die Prüfung der Hypothesen H1 und H2 erfolgte über zweifaktorielle Varianzanalysen (ANOVA) mit der jeweiligen abhängigen Variable als Kriterium. Als Faktoren wurden die CS-Präsenz (HS vs. HS + CS) sowie die Teilnehmendengruppe (TG-M vs. TG-eM) berücksichtigt. Für die beiden Indikatoren stereotypbezogener Zuschreibungen (Eigenschaftszuschreibung und Distanz/Vermeidung) wurden zusätzlich multivariate Varianzanalysen (MANOVA) berechnet, gefolgt von univariaten Analysen zur differenzierten Betrachtung der Einzeleffekte.

Die explorative Forschungsfrage EF1 zur Interventionsbereitschaft wurde über separate zweifaktorielle Varianzanalysen für die drei Dimensionen (Schweregradwahrnehmung, Verantwortungsgefühl, Interventionsabsicht) untersucht.

Effekte der CS-Quelle (H1a, H2a) wurden in einer separaten Analyse innerhalb der Teilstichprobe mit Counter Speech geprüft. Hierzu wurden zweifaktorielle Varianzanalysen mit der CS-Quelle (CS-M vs. CS-eM) und den Teilnehmendengruppen (TG-M vs. TG-eM) als Faktoren berechnet.

Ergänzend zu den hypothesengeleiteten Analysen wurden explorative Zusatzanalysen durchgeführt, um zu prüfen, ob sich zentrale Effekte durch ausgewählte individuelle Personenmerkmale verändern. Berücksichtigt wurden politische Orientierung, vorbestehende Einstellungen gegenüber ethnischen Minderheiten sowie Social-Media-Nutzungsvariablen. Weitere erhobene Merkmale wurden nicht systematisch als Moderatoren in allen explorativen Modellen integriert, da dies die Modellkomplexität und die Anzahl explorativer Tests deutlich

erhöht hätte. Die Zusatzanalysen beschränkten sich zudem auf die zentralen hypothesengeleiteten Wirkungszusammenhänge der Studie (H1, H2 und EF1).

5.7 Forschungsethische Aspekte

Die Untersuchung von Hate Speech und Counter Speech betraf sensible Themen wie Rassismus, gruppenbezogene Abwertung und potenziell belastende Inhalte. Daher wurden für die vorliegende Studie zentrale forschungsethische Standards berücksichtigt, um Risiken für Teilnehmende zu minimieren und eine verantwortungsvolle Durchführung sicherzustellen (Department of Communication, 2019).

Ein zentrales Prinzip war *Minimizing Risk of Harm*. Da die Teilnehmenden im Experiment Hate Speech in einem realitätsnahen Kommentarsetting rezipieren, wurde bereits im Briefing transparent auf potenziell belastende Inhalte hingewiesen. Die Teilnahme erfolgte ausschließlich freiwillig und erst nach einer informierten Einwilligung. Zudem hatten die Teilnehmenden jederzeit die Möglichkeit, den Fragebogen ohne Angabe von Gründen abzubrechen.

Zum Schutz potenziell vulnerabler Gruppen wurden nur Personen ab 18 Jahren in die Befragung einbezogen. Ferner wurden die Stimuli realitätsnah gestaltet und als fiktive Inhalte gekennzeichnet. Nutzernamen und Profilbilder wurden neutralisiert bzw. anonymisiert, um keine realen Personen abzubilden und unbeabsichtigte Re-Identifizierbarkeit oder Stigmatisierung zu vermeiden.

Ein weiteres Kernprinzip betraf Anonymität und Vertraulichkeit (*protecting anonymity and confidentiality*) (Department of Communication, 2019). Die Erhebung erfolgte anonym über einen standardisierten Onlinefragebogen. Es wurden keine direkt identifizierenden personenbezogenen Daten (z. B. Name, E-Mail-Adresse, IP-Adresse) erhoben. Die Speicherung und Auswertung der Daten erfolgten ausschließlich zu wissenschaftlichen Zwecken und unter Beachtung der geltenden datenschutzrechtlichen Vorgaben.

Nach Abschluss der Befragung erhielten alle Teilnehmenden ein ausführliches Debriefing, in dem Zielsetzung und Aufbau der Studie erläutert sowie die Inhalte als fiktiv ausgewiesen wurden. Zusätzlich wurden Kontakt- und Unterstützungsstellen angegeben, um bei möglichen negativen Emotionen oder Bedarf an Beratung niedrigschwellige Hilfsangebote zugänglich zu machen.

6 Ergebnisse

Dieses Kapitel berichtet über die Ergebnisse der experimentellen Studie. Im Anschluss an die Stichprobenbereinigung werden Randomisierung, Skalenqualität, deskriptive Kennwerte sowie die Hypothesentests zur wahrgenommenen sozialen Identitätsbedrohung (H1/H1a) und stereotypbezogenen Zuschreibungen (H2/H2a) sowie die explorative Forschungsfrage zur Interventionsbereitschaft (EF1) dargestellt. Ergänzend werden explorative Zusatzanalysen zur Einordnung potenzieller Moderatoren berichtet.

Alle Analysen wurden in IBM SPSS Statistics durchgeführt; Effektstärken werden ergänzend berichtet (partielles η^2, Cohen's d).

6.1 Analysestichprobe

Der Rohdatensatz wurde in IBM SPSS Statistics importiert und aufbereitet. Variablen- und Wertelabels wurden entsprechend dem Codebuch vergeben, Messniveaus konsistent definiert und fehlende Werte als benutzerdefinierte Missing-Codes (-9) gekennzeichnet. Da alle Items als Pflichtfragen implementiert waren, traten keine systematischen Item-Missings auf. Skalenwerte wurden daher als Mittelwertindizes über die jeweiligen Items gebildet.

Die Stichprobenbereinigung reduzierte den Ausgangsdatensatz ($N = 265$) auf eine bereinigte Analysestichprobe von $N = 191$. Für Analysen mit den Teilnehmendengruppen wurden Fälle ohne Angabe zur ethnischen Selbstidentifikation ausgeschlossen ($n = 4$), sodass die Hypothesentests auf $N = 187$ komplett ausgefüllten Fragebögen basierten. Der Bereinigungsprozess ist in Tabelle 3 dokumentiert.

Tabelle 3

Stichprobenbereinigung und Ausschlusskriterien

Ausschlusskriterium / Qualitätsprüfung	**Ausgeschlossen (*n*)**	**Verbleibend (*n*)**
Rohdatensatz (Teilnahmen insgesamt)	–	265
Abgebrochene Fragebögen (unvollständige Teilnahme)	57	208
Kein Einverständnis und/oder fehlende Bestätigung der Volljährigkeit	3	205
Aufmerksamkeitskontrolle zum Stimulus falsch beantwortet	12	193
Formale Einschlusskriterien nicht erfüllt (Minderjährigkeit nach Altersangabe bzw. Wohnort außerhalb des Zielraums)	2	191
Straightlining-Prüfung (kein Ausschluss, nur Dokumentation)	0	191
Speeding-Prüfung (kein Ausschluss, nur Dokumentation)	0	191
Keine Angabe zur ethnischen Selbstidentifikation (Ausschluss in Analysen mit Teilnehmendengruppen)	4	187

Ergänzend wurden Hinweise auf potenziell unaufmerksames Antwortverhalten geprüft. Dazu wurde das Antwortmuster zunächst hinsichtlich Straightlining untersucht. Ein rein technischer Straightlining-Indikator wurde allerdings nicht als automatisches

Ausschlusskriterium verwendet, da gleichförmige Antwortmuster bei konsistent formulierten Likert-Skalen auch stabile Einstellungen widerspiegeln konnten (Zhang & Conrad, 2014, S. 127ff.). Zur weiteren Plausibilisierung erfolgte eine inhaltliche Konsistenzprüfung, bei der positive Stereotyp-Items (D102_01–D102_04) den gegensätzlich gepolten Distanz-/Vermeidungs-Items (D103_01–D103_03) gegenübergestellt wurden. Hierbei zeigten sich keine Auffälligkeiten.

Zusätzlich wurden Bearbeitungszeiten als Speeding-Indikator berücksichtigt. Als Richtwert galt eine Mindestbearbeitungszeit von .3 Sekunden pro Wort (Sensitivitätsanalyse: .25 Sekunden pro Wort), wobei textreiche Einstiegsseiten (z. B. Begrüßung, Datenschutzhinweise) von der Berechnung ausgenommen wurden. Da auffällige Bearbeitungszeiten nicht systematisch mit fehlerhaften Antworten in der Aufmerksamkeitskontrolle oder weiteren Qualitätsindikatoren einhergingen, wurde Speeding nicht als Ausschlusskriterium verwendet, sondern lediglich dokumentiert (Zhang & Conrad, 2014, S. 127ff.).

Für die Hypothesentests wurden die in Kapitel 5 beschriebenen Auswertungsgruppen verwendet (CS-Präsenz und CS-Quelle). Die jeweiligen Zellgrößen der Experimentalbedingungen und Analysegruppen sind Tabelle 4 zu entnehmen.

Tabelle 4

Zellgrößen der Experimentalbedingungen und Analysegruppen

Analyseebene	Gesamtfälle	Bedingung	*n*	%	Relevante Analysen
Experimentalbedingungen (bereinigte Stichprobe)	191	HS	66	34.6	Randomisierungscheck
	191	HS + CS-M	63	33.0	Randomisierungscheck
	191	HS + CS-eM	62	32.5	Randomisierungscheck
Vergleich CS-Präsenz	187	HS	63	33.7	H1, H2, EF1
	187	HS + CS	124	66.3	H1, H2, EF1
Vergleich der CS-Quelle	124	CS-M	63	50.8	H1a, H2a
	124	CS-eM	61	49.2	H1a, H2a

Anmerkung. HS = Hate Speech; CS = Counter Speech; CS-M = Counter Speech aus der Mehrheitsgesellschaft; CS-eM = Counter Speech aus der ethnischen Minderheit.

6.2 Randomisierung und Skalenqualität

6.2.1 Randomisierung

Zur Prüfung der erfolgreichen Randomisierung wurden die drei Experimentalbedingungen (HS, HS + CS-M, HS + CS-eM) hinsichtlich zentraler soziodemografischer Merkmale verglichen ($N = 191$). Für Alter wurde eine einfaktorielle ANOVA berechnet, kategoriale Variablen (Geschlecht, Wohnort, Bildungsabschluss, ethnische Selbstidentifikation) wurden mittels Chi-Quadrat-Tests geprüft.

Zwischen den drei Gruppen zeigten sich keine signifikanten Unterschiede im Alter, $F(2, 188) = .675$, $p = .510$ (Welch: $p = .548$). Zudem gab es keine signifikanten Unterschiede für Geschlecht, Wohnort und ethnische Selbstidentifikation (alle $p > .05$). Für den Bildungsabschluss zeigte sich ein signifikanter Gruppenunterschied, $\chi^2(10) = 19.918$, $p = .030$, Cramér's V = .228. Da in mehreren Zellen erwartete Häufigkeiten < 5 auftraten, wurde dieser Befund vorsichtig interpretiert.

Zusätzlich wurde der für die Hypothesentests zentrale Kontrast HS vs. HS + CS hinsichtlich soziodemografischer Variablen geprüft ($N = 187$; Ausschluss fehlender Angaben zur ethnischen Selbstidentifikation). Für Alter, Wohnort, Bildungsabschluss und ethnische Selbstidentifikation zeigten sich keine signifikanten Unterschiede zwischen den Bedingungen (alle $p > .05$). Für das Geschlecht ergab sich allerdings ein signifikanter Unterschied, $\chi^2(2) = 6.68$, $p = .035$, Cramér's V = .19, wobei in mehreren Zellen erwartete Häufigkeiten unter 5 lagen. Dieser Befund wird entsprechend vorsichtig interpretiert.

Innerhalb der Counter-Speech-Teilstichprobe ($N = 124$) wurden zudem die beiden CS-Quellen (CS-M vs. CS-eM) verglichen. Dabei zeigten sich keine Unterschiede für Wohnort und ethnische Selbstidentifikation (alle $p > .05$). Es gab einen signifikanten Unterschied im Bildungsabschluss, $\chi^2(5) = 12.841$, $p = .025$, V = .322, der ebenfalls aufgrund kleiner erwarteter Zellhäufigkeiten eingeschränkt interpretierbar war.

Die Randomisierung war weitgehend erfolgreich, allerdings zeigten sich vereinzelte Unterschiede in Bildungsabschluss und Geschlecht, die bei der Interpretation berücksichtigt werden müssen. Die vollständigen Randomisierungstests sind im Anhang (siehe Tabelle D1) dokumentiert.

6.2.2 Skalenqualität

Zur Beurteilung der Messqualität wurde für alle mehrdimensionalen Skalen die interne Konsistenz mittels Cronbach's α berechnet. Als Orientierungswerte wurden die von George und Mallery (2019) vorgeschlagenen Grenzwerte herangezogen ($\alpha > .90$ exzellent; > .80 gut; > .70 akzeptabel; > .60 fragwürdig; > .50 niedrig; < .50 inakzeptabel).

Die Skala zur wahrgenommenen sozialen Identitätsbedrohung (5 Items) weist mit $\alpha = .902$ eine sehr hohe interne Konsistenz auf. Auch die beiden Indikatoren der stereotypbezogenen Zuschreibung zeigen gute bis sehr gute Reliabilitätswerte. Eigenschaftszuschreibung (4 Items), $\alpha = .860$, sowie Distanz/Vermeidung (3 Items), $\alpha = .893$. Entsprechend der Operationalisierung wurden die Dimensionen getrennt ausgewertet. Explorative Gesamtindizes werden nur ergänzend berichtet.

Die Interventionsbereitschaft wurde entsprechend der Prozesslogik getrennt ausgewertet. Schweregradwahrnehmung (3 Items; α = .845), Verantwortungsgefühl (3 Items; α = .814) sowie Interventionsabsicht (Einzelitem). Hier ergab eine explorative Gesamtskala zwar eine gute Reliabilität (α = .820), wurde dennoch getrennt analysiert, da die unterscheidbaren Prozessschritte abgebildet werden und somit die Prozesslogik verdeckt werden könnte.

Für die Mediennutzung wurde ebenfalls ein zusammenfassender Index aus drei Items geprüft. Dieser zeigte jedoch eine unzureichende interne Konsistenz (α = .455), weshalb auf die Bildung einer Gesamtskala verzichtet wurde. Daher wurden die Mediennutzungsindikatoren in späteren Zusatzanalysen nur als Einzelvariablen berücksichtigt.

Eine Übersicht über alle Reliabilitätskennwerte ist der Tabelle 5 zu entnehmen.

Tabelle 5

Interne Konsistenzen der Skalen (Cronbach's α)

Skala	Items	Cronbach's α
Wahrgenommene soziale Identitätsbedrohung (D101_01–D101_05)	5	.902
Stereotypbezogene Zuschreibung – Eigenschaftszuschreibung (D102_01–D102_04)	4	.860
Stereotypbezogene Zuschreibung – Distanz/Vermeidung (D103_01–D103_03)	3	.893
Explorative Gesamt-stereotypbezogene Zuschreibung (D102 + D103)	7	.826
Interventionsbereitschaft – Schweregradwahrnehmung (D104_01–D104_03)	3	.845
Interventionsbereitschaft – Verantwortungsgefühl (D105_01–D105_03)	3	.814
Interventionsbereitschaft – Interventionsabsicht (D106_01)	1	_
Explorative Gesamt-Interventionsbereitschaft (D104 + D105 + D106)	7	.820
Mediennutzung (B103_01–B103_03)	3	.455

6.3 Deskriptive Statistik der zentralen Variablen

6.3.1 Stichprobenbeschreibung

Die bereinigte Stichprobe (N = 191) hatte ein durchschnittliches Alter von M = 34.2408 Jahren (Median = 28; Range: 18–84). Ein Anteil von 64.9 % (n = 124) identifizierten sich als weiblich, 34.6 % (n = 66) als männlich, .5 % (n = 1) machten keine Angabe. 61.8 % (n = 118) lebten in Deutschland, 38.2 % (n = 73) in Österreich, für die Schweiz wurden keine Angaben gemacht. 19.9 % (n = 38) ordneten sich einer ethnischen Minderheit zu, 78.0 % (n = 149) der Mehrheitsgesellschaft, 2.1 % (n = 4) machten keine Angabe. Das Bildungsniveau ist insgesamt hoch. Mittlere Reife/Sekundärschule: 8.9 % (n = 17), Abitur/Matura: 19.9 % (n = 38), Bachelor-Abschluss /Diplom /FH-Diplom: 45.5 % (n = 87), Master/Magister: 22.5 % (n = 43), Promotion: 1.0 % (n = 2), Andere: 2.1 % (n = 4).

6.3.2 Deskriptive Kennwerte der zentralen Variablen

Für alle Skalen wurden Mittelwerte gebildet. Tabelle 6 zeigt Mittelwert, Standardabweichung sowie die Verteilungskennwerte.

Wahrgenommene soziale Identitätsbedrohung liegt im Mittel im niedrigen bis moderaten Bereich (*M* = 2.3435; SD = 1.09301). Die stereotypbezogene Zuschreibung (Eigenschaftszuschreibung) gegenüber ethnischen Minderheiten fällt insgesamt eher positiv aus (M = 3.5118; SD = .77599), während stereotypbezogene Zuschreibung (Distanz/Vermeidung) im Mittel im unteren Bereich lag (M = 1.4101; *SD* = .67098) und eine ausgeprägte Bodenverteilung zeigt. Aufgrund deutlicher Abweichungen von der Normalverteilung wurden die Ergebnisse hier vorsichtig interpretiert. In Bezug auf die Interventionsbereitschaft wird die Schweregradwahrnehmung des Kommentars hoch bewertet (*M* = 4.2478; *SD* = .8195), das Verantwortungsgefühl liegt im mittleren Bereich (*M* = 3.3333; *SD* = 1.01394), während die Interventionsabsicht (Einzelitem) eher niedrig ausfällt (*M* = 2.2300; *SD* = 1.2650).

Tabelle 6

Deskriptive Kennwerte der zentralen Variablen (N =191)

Skala	***M***	***SD***	**Schiefe**	**Min – Max**	**Kurtosis**
Wahrgenommene soziale Identitätsbedrohung	2.3435	1.09301	.497	1 bis 5	-.544
Stereotypbezogene Zuschreibung – Eigenschaftszuschreibung	3.5118	.77599	-.327	1 bis 5	.872
Stereotypbezogene Zuschreibung – Distanz/Vermeidung	1.4101	.67098	2.208	1 bis 5	5.792
Interventionsbereitschaft – Schweregradwahrnehmung	4.2478	.81951	-1.403	1 bis 5	2.270
Interventionsbereitschaft – Verantwortungsgefühl	3.3333	1.01394	-.286	1 bis 5	-.594
Interventionsbereitschaft – Interventionsabsicht (Einzelitem)	2.2300	1.26500	.693	1 bis 5	-.708

6.3.3 Deskriptive Beschreibung zentraler soziodemografischer- und Moderatorvariablen

Zur Einordnung der Hauptergebnisse wurden wichtige soziodemografische- und Moderatorvariablen deskriptiv ausgewertet (für eine Übersicht siehe Tabelle 7). Berichtet wurden Voreinstellungen gegenüber ethnischen Minderheiten, politische Orientierung sowie Social-Media-Nutzung (Nutzungsdauer und Nutzungsformen).

Die allgemeine Voreinstellung gegenüber ethnischen Minderheiten wurde auf einer Skala von -5 (sehr negativ) bis +5 (sehr positiv) erhoben. Insgesamt zeigt die Stichprobe eine überwiegend positive Grundhaltung. Der Mittelwert liegt bei *M* = 2.5759 (*SD* = 1.81608), der Median bei *Md* = 3.00. Die Werte reichen von -3 bis 5. Die Verteilung ist leicht linksschief (Schiefe = -.635), was auf eine Häufung positiver Einstellungswerte hinweist. Zur zusätzlichen inhaltlichen Einordnung wurde eine dreistufige Kategorienvariable gebildet: 86.9

% der Befragten zeigten eine positive Voreinstellung (+1 bis +5), 5.2 % ordneten sich neutral (0) ein und 7.9 % berichteten von einer negativen Voreinstellung (-5 bis -1). Dieses Muster weist auf eine insgesamt eher minderheitenfreundliche Stichprobe hin.

Die politische Orientierung wurde über eine Selbstverortung auf einer Links-Rechts-Skala erfasst und für die Auswertung auf einen Wertebereich von 0 bis 100 transformiert. Die Stichprobe ist insgesamt eher linkszentriert: Der Mittelwert beträgt *M* = 30.25 (*SD* = 23.291), der Median *Md* = 24.00. Der Wertebereich reicht von 1 bis 100, die Verteilung ist deutlich rechtsschief (Schiefe = 1.002), was auf eine Häufung niedriger (linker) Werte und einzelne höhere Ausprägungen hinweist. Ergänzend wurde eine explorative Dreiteilung vorgenommen. 64.9 % wurden als „links“ (0-33), 26.2 % als „Mitte“ (34-66) und 8.9 % als „rechts“ (67-100) klassifiziert. Diese Einordnung dient ausschließlich der Kontextualisierung und wird nicht als inferenzstatistischer Test interpretiert.

Das Social-Media-Nutzungsverhalten wurde über drei Indikatoren (Informationssuche, Kommentieren, Teilen) sowie über die tägliche Nutzungsdauer erfasst (jeweils 5-stufige Skalen). Insgesamt zeigt sich ein überwiegend konsumorientiertes Nutzungsprofil. Informationssuche wird häufig genutzt (*M* = 3.66; *SD* = 1.078; *Md* = 4; Modus = 4). 61.8 % gaben an, soziale Medien oft oder sehr oft zur Informationssuche zu verwenden (38,2 % „oft“, 23.6 % „sehr oft“), während 38.2 % dies nur selten bis manchmal tun. Das Kommentieren von Beiträgen ist dagegen selten (*M* = 1.62; *SD* = .737; *Md* = 1). 50.8 % kommentieren nie und weitere 38.7 % selten; 8.9 % manchmal, oft 1% und sehr oft .5%. Teilen von Inhalten liegt im mittleren Bereich (*M* = 2.66; *SD* = 1.107; *Md* = 3; Modus = 3). Nie 16.2%, selten 29.3%, 32.5 % teilen manchmal Inhalte, 16.2 % oft und 5.8 % sehr oft. Die Social-Media-Nutzungsintensität zeigt überwiegend moderate Nutzung. Im Mittel liegt die Kategorie bei *M* = 2.38 (*SD* = 1.043). Am häufigsten wurden 1 bis 2 Stunden pro Tag berichtet (34.0 %), gefolgt von 2 bis 3 Stunden (29.8 %) und unter 1 Stunde (22.5 %). Intensivere Nutzung (≥ 3 Stunden) ist vergleichsweise selten (13.6 %).

Tabelle 7

Deskriptive Kennwerte der Moderatorvariablen (N = 191)

Variable	*M*	*SD*	Min–Max	Schiefe
Voreinstellung ggü. ethnischen Minderheiten	2.5759	1.81608	-3 bis 5	-.635
Politische Orientierung	30.25	23.291	1 bis 100	1.002
Mediennutzung – Informationen suchen	3.66	1.078	1 bis 5	-.585
Mediennutzung – Beiträge kommentieren	1.62	.737	1 bis 5	1.223
Mediennutzung – Informationen teilen	2.66	1.107	1 bis 5	.236
Social-Media-Nutzungsintensität	2.38	1.043	1 bis 5	.430

6.4 Hypothesentests und explorative Forschungsfrage

Die Hypothesentests basierten auf $N = 187$ Fällen (TG-eM: $n = 38$; TG-M: $n = 149$). Analysen zur CS-Quelle wurden innerhalb der Counter-Speech-Teilstichprobe ($N = 124$) durchgeführt. Aufgrund kleiner Zellgrößen in TG-eM wurden Ergebnisse zur CS-Quelle vorsichtig interpretiert.

6.4.1 H1: Wahrgenommene soziale Identitätsbedrohung

Zur Prüfung von H1 wurde eine zweifaktorielle Varianzanalyse (2 × 2 ANOVA) mit wahrgenommener soziale Identitätsbedrohung als abhängiger Variable sowie CS-Präsenz (HS vs. HS + CS) und Teilnehmendengruppe (TG-M vs. TG-eM) als Faktoren berechnet.

Es zeigte sich ein signifikanter Haupteffekt der CS-Präsenz, $F(1, 183) = 6.04$, $p = .015$, $\eta p^2 = .032$. Dieser Haupteffekt wird allerdings durch die signifikante Interaktion zwischen CS-Präsenz und Teilnehmendengruppe qualifiziert, $F(1, 183) = 6.22$, $p = .013$, $\eta p^2 = .033$. In den einfachen Effekten reduzierte Counter Speech die wahrgenommene soziale Identitätsbedrohung ausschließlich bei Teilnehmenden ethnischer Minderheiten, hingegen nicht in der Mehrheitsgesellschaft.

Damit findet Hypothese H1 teilweise empirische Unterstützung. Die wahrgenommene soziale Identitätsbedrohung war unter HS + CS geringer als unter HS, und der Effekt trat ausschließlich in TG-eM auf. Der Interaktionseffekt sollte aufgrund kleiner Zellgrößen insbesondere in der Teilnehmendengruppe (TG-eM) vorsichtig interpretiert werden. Deskriptive Werte, Varianzanalysen und einfache Effekte sind Tabelle 8 zu entnehmen.

Tabelle 8

Wahrgenommene sozialen Identitätsbedrohung nach CS-Präsenz und Teilnehmendengruppen

Gruppe	HS: *M*	HS: *SD*	HS: *n*	HS + CS: *M*	HS + CS: *SD*	HS + CS: *n*	Gesamt: *M*	Gesamt: *SD*	Gesamt: *n*
TG-eM	3.5529	1.11025	17	2.6095	1.09449	21	3.0316	1.18598	38
TG-M	2.1870	1.03056	46	2.1942	.99625	103	2.1919	1.00347	149
Gesamt	2.5556	1.20934	63	2.2645	1.02093	124	2.3626	1.09352	187

Panel A *Varianzanalytische Effekte (2×2 ANOVA)*

Effekt	F(df1, df2)	*p*	ηp²
CS-Präsenz (HS vs. HS + CS)	$F(1, 183) = 6.036$	.015	.032
Teilnehmendengruppen (TG-M vs. TG-eM)	$F(1, 183) = 21.853$	< .001	.107
Interaktion	$F(1, 183) = 6.224$	.013	.033

Panel B *Einfache Effekte innerhalb der Teilnehmendengruppen*

Gruppe	ΔM (HS, HS + CS)	*p*	95%-KI Untergrenze	95%-KI Obergrenze
TG-eM	.943	.005	.283	1.604
TG-M	-.007	.968	-.366	.352

Anmerkung. HS = Hate Speech; CS = Counter Speech; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Berichtet werden Zellenmittelwerte und Varianzanalysen einer zweifaktoriellen Varianzanalyse (2 × 2 ANOVA) mit partieller Effektstärke ηp². Δ*M* = Mittelwertdifferenz (HS minus HS + CS); KI = Konfidenzintervall.

6.4.2 H1a: CS-Quelle und wahrgenommene soziale Identitätsbedrohung

Zur Prüfung der H1a wurde innerhalb der HS + CS-Teilstichprobe ($N = 124$) eine zweifaktorielle Varianzanalyse durchgeführt. Die abhängige Variable war die wahrgenommener sozialer Identitätsbedrohung und die Faktoren waren CS-Quelle (CS-M vs. CS-eM) und Teilnehmendengruppe (TG-M vs. TG-eM) (Levene: $p = .891$).

Es zeigte sich weder ein signifikanter Haupteffekt der CS-Quelle, $F(1, 120) = .98$, $p = .325$, $\eta p^2 = .008$, noch eine signifikante Interaktion zwischen CS-Quelle und Teilnehmendengruppe, $F(1, 120) = 1.01$, $p = .318$, $\eta p^2 = .008$. Ein Haupteffekt der Teilnehmendengruppe zeigte sich lediglich tendenziell, $F(1, 120) = 3.37$, $p = .069$, $\eta p^2 = .027$.

Deskriptiv zeigte sich für die Teilnehmendengruppe der ethnischen Minderheit eine höhere Bedrohung bei Counter Speech aus der ethnischen Minderheit ($M = 2.8889$, $SD = .16237$) im Vergleich zu Counter Speech aus der Mehrheitsgesellschaft ($M = 2.4000$, $SD = 1.04098$). Der Unterschied war nicht signifikant.

Damit findet Hypothese H1a keine empirische Unterstützung. Deskriptive Statistiken und ANOVA-Ergebnisse sind Tabelle 9 zu entnehmen.

Tabelle 9

Wahrgenommene soziale Identitätsbedrohung nach CS-Quelle und Teilnehmendengruppen

Gruppe	CS-M			CS-eM		
	M	***SD***	***n***	***M***	***SD***	***n***
TG-eM	2.4000	1.04098	12	2.8889	1.16237	9
TG-M	2.1961	1.00597	51	2.1923	.99643	52
Gesamt	2.2349	1.00742	63	2.2951	1.04218	61

Panel A *Varianzanalytische Effekte (2 × 2 ANOVA)*

Effekt	F(df1, df2)	*p*	ηp²
CS-Quelle	$F(1, 120) = .977$	.325	.008
Teilnehmendengruppen	$F(1, 120) = 3.365$	.069	.027
Interaktion (CS-Quelle × TG)	$F(1, 120) = 1.007$	.318	.008
Levene-Test (Varianzhomogenität)	*p*	.891	–

Anmerkung. CS-M = Counter Speech aus der Mehrheitsgesellschaft; CS-eM = Counter Speech aus der ethnischen Minderheit; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Analysen basieren auf der HS + CS-Teilstichprobe (N = 124). Berichtet werden Varianzanalysen einer zweifaktoriellen Varianzanalyse (2 × 2 ANOVA) mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

6.4.3 H2: Stereotypbezogene Zuschreibungen (Eigenschaftszuschreibung und Distanz/Vermeidung)

Zur Prüfung von H2 wurde eine MANOVA mit Eigenschaftszuschreibung und Distanz/Vermeidung als abhängigen Variablen sowie CS-Präsenz (HS vs. HS + CS) und Teilnehmendengruppe (TG-M vs. TG-eM) als Faktoren berechnet. Als multivariates Testkriterium wird die Pillai-Spur berichtet, da sie als robust gegenüber Annahmeverletzungen gilt.

Es ergaben sich keine signifikanten multivariaten Effekte für die CS-Präsenz (Pillai-Spur = .020; $F(2, 182) = 1.877$, $p = .156$, $\eta p^2 = .020$), für die Teilnehmendengruppen (Pillai-Spur = .013; $F(2, 182) = 1.159$, $p = .316$, $\eta p^2 = .013$) sowie für die Interaktion (Pillai-Spur = .017; $F(2, 182) = 1.535$, $p = .218$, $\eta p^2 = .017$).

Auch in den univariaten Folgeanalysen zeigten sich keine signifikanten Effekte. Für die Eigenschaftszuschreibung ergaben sich weder für die CS-Präsenz ($p = .140$), noch für die Teilnehmendengruppen ($p = .221$) oder die Interaktion ($p = .161$) signifikante Unterschiede. Für die Variable Distanz/Vermeidung ergaben sich ebenfalls keine signifikanten Effekte (CS-Präsenz: $p = .454$; Teilnehmendengruppen: $p = .621$; Interaktion: $p = .563$).

Damit wird Hypothese H2 nicht unterstützt. Deskriptive Werte, multivariate Tests und univariate Folgeanalysen sind Tabelle 10 zu entnehmen.

Tabelle 10

Stereotypbezogene Zuschreibung nach CS-Präsenz und Teilnehmendengruppen

Variable	HS-eM *M*	HS- eM *SD*	*n*	HS-M *M*	HS-M *SD*	n	HS + CS-eM *M*	HS + CS-eM *SD*	n	HS + CS-M *M*	HS + CS-M *SD*	n
Eigenschaftszuschreibung	3.8676	.63195	17	3.4891	.82151	46	3.4524	.83524	21	3.4782	.75983	103
Distanz/Vermeidung	1.5490	1.18989	17	1.4130	.59290	46	1.3810	.74748	21	1.3916	.58750	103

Panel A *Multivariate Tests (Pillai-Spur)*

Effekt	Pillai-Spur	*F*	*p*	ηp²
CS-Präsenz (HS vs. HS + CS)	.020	*F*(2, 182)=1.877	.156	.020
Teilnehmendengruppen (TG-M vs. TGeM)	.013	*F*(2, 182)=1.159	.316	.013
Interaktion	.017	*F*(2, 182)=1.535	.218	.017

Panel B *Univariate Folgeanalysen*

Abhängige Variable	Effekt	*F*	*p*	ηp²
Eigenschaftszuschreibung	CS-Präsenz	*F*(1, 183)=2.200	.140	.012
Eigenschaftszuschreibung	Teilnehmendengruppen	*F*(1, 183)=1.507	.221	.008
Eigenschaftszuschreibung	Interaktion	*F*(1, 183)=1.979	.161	.011
Distanz/Vermeidung	CS-Präsenz	*F*(1, 183)= .562	.454	.003
Distanz/Vermeidung	Teilnehmendengruppen	*F*(1, 183)= .246	.621	.001
Distanz/Vermeidung	Interaktion	*F*(1, 183)= .336	.563	.002

Anmerkung. HS = Hate Speech; CS = Counter Speech; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Multivariate Ergebnisse basieren auf einer MANOVA (Panel A; Testkriterium: Pillai-Spur) und univariate Folgeanalysen auf zweifaktoriellen Varianzanalysen (2 × 2 ANOVA; Panel B) mit partieller Effektstärke ηp².

6.4.4 H2a: CS-Quelle und stereotypbezogene Zuschreibungen

Zur Prüfung von H2a wurde innerhalb der Bedingung HS + CS ($N = 124$) eine zweifaktorielle Varianzanalyse mit der CS-Quelle (CS-M vs. CS-eM) und den Teilnehmendengruppen (TG-M vs. TG-eM) durchgeführt. Als abhängige Variablen dienten wieder Eigenschaftszuschreibung und Distanz/Vermeidung.

Es zeigten sich weder für Eigenschaftszuschreibung signifikante Effekte der CS-Quelle ($p = .086$) oder der Interaktion ($p = .184$) noch für Distanz/Vermeidung (CS-Quelle: $p = .111$; Interaktion: $p = .976$).

In einer explorativen Subgruppenanalyse innerhalb der Mehrheitsgesellschaft ergab sich für Distanz/Vermeidung ein signifikanter Unterschied zugunsten von Counter Speech aus der ethnischen Minderheit (Welch-t-Test wegen Varianzenungleichheit; Levene $p = .008$; $t(89.44) = -2.04$, $p = .044$). Dieser Effekt wird im Gesamtmodell nicht bestätigt (zweifaktorielle ANOVA: CS-Quelle $F(1,120) = 2.571$, $p = .111$; Interaktion $F(1,120) = .001$, $p = .976$) und zeigt sich auch in den einfachen Effekten innerhalb der ANOVA nur als Trend

($F = 3.717$, $p = .056$). Dieser Befund wird daher aufgrund multipler univariater Tests und der Abhängigkeit vom Auswertungsweg (t-Test vs. Modelltest) als explorativ berichtet.

Damit finden sich keine signifikanten Effekte im Sinne von Hypothese H2a. Deskriptive Kennwerte und Varianzanalysen sind Tabelle 11 zu entnehmen.

Tabelle 11

Stereotypbezogene Zuschreibung nach CS-Quelle und Teilnehmendengruppen

CS-Quelle	Teilnehmenden-gruppen	N	Eigenschaftszuschreibung M	SD	Distanz/ Vermeidung M	SD
CS-M	TG-eM	12	3.2083	.57241	1.2778	.66414
CS-M	TG-M	51	3.4412	.81331	1.2745	.45561
CS-eM	TG-eM	9	3.7778	1.04167	1.5185	.86781
CS-eM	TG-M	52	3.5144	.70955	1.5064	.67798
Gesamt	TG-eM	21	3.4524	.83524	1.3810	.74748
Gesamt	TG-M	103	3.4782	.75983	1.3916	.58750
Gesamt	Gesamt	124	3.4738	.76961	1.3898	.61408

Panel A *Tests der Zwischensubjekteffekte (zweifaktorielle ANOVA)*

Abhängige Variable	Effekt	*F*	df1	df2	*p*	ηp²	Levene *p*
Eigenschaftszuschreibung	CS-Quelle	2.99	1	120	.086	.024	.196
	Teilnehmendengruppen	.007	1	120	.935	.000	
	Interaktion	1.78	1	120	.184	.015	
Distanz/Vermeidung	CS-Quelle	2.57	1	120	.111	.021	.022
	Teilnehmendengruppen	.003	1	120	.958	.000	
	Interaktion	.001	1	120	.976	.000	

Anmerkung. CS-M = Counter Speech aus der Mehrheitsgesellschaft; CS-eM = Counter Speech aus der ethnischen Minderheit; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Berichtet werden Tests der Zwischensubjekteffekte aus zweifaktoriellen Varianzanalysen (2 × 2 ANOVA) mit partieller Effektstärke ηp². Levene-Tests prüfen Varianzhomogenität.

6.4.5 EF1: Interventionsbereitschaft

Zur Prüfung von EF1 wurden drei 2 × 2 ANOVAs mit CS-Präsenz und Teilnehmendengruppe als Faktoren berechnet. Abhängige Variablen waren Schweregradwahrnehmung, Verantwortungsgefühl und Interventionsabsicht ($N = 187$).

Für die Schweregradwahrnehmung ergab sich kein Haupteffekt der CS-Präsenz ($p = .435$), allerdings ein signifikanter Haupteffekt der Teilnehmendengruppen, $F(1, 183) = 4.203$, $p = .042$, $\eta p^2 = .022$. Die Teilnehmendengruppe der Mehrheitsgesellschaft bewertete die Schweregradwahrnehmung ($M = 4.3333$, SD = .78461) höher als die Teilnehmendengruppe der ethnischen Minderheit ($M = 3.9825$, $SD = .82363$). Die Interaktion war nicht signifikant ($p = .065$). Für Verantwortungszuschreibung zeigten sich keine signifikanten Effekte (CS-Präsenz: $p = .794$; Teilnehmendengruppen: $p = .916$; Interaktion: $p = .843$). Ebenso ergaben

sich für Interventionsabsicht keine signifikanten Effekte (CS-Präsenz: $p = .179$; Teilnehmendengruppen: $p = .186$; Interaktion: $p = .181$).

Für Verantwortungsgefühl und Interventionsabsicht ergaben sich keine signifikanten Effekte der CS-Präsenz oder der Interaktion. Ein Gruppenunterschied zeigte sich ausschließlich für die Schweregradwahrnehmung (TG-M > TG-eM). Zellmittelwerte und Varianzanalysen aller drei Variablen sind den Tabellen 12, 13 und 14 zu entnehmen.

Tabelle 12

Schweregradwahrnehmung nach CS-Präsenz und Teilnehmendengruppen

Bedingung	Teilnehmendengruppen	*n*	*M*	*SD*	Gesamt *n*	Gesamt *M*	Gesamt *SD*	Levene *p*
HS	TG-eM	17	4.20	.96	63	4.22	.88	.345
HS	TG-M	46	4.22	.86	63	4.22	.88	.345
HS + CS	TG-eM	21	3.81	.67	124	4.28	.76	.345
HS + CS	TG-M	103	4.38	.75	124	4.28	.76	.345
Gesamt	TG-eM	38	3.98	.82	187	4.26	.80	.345
Gesamt	TG-M	149	4.33	.78	187	4.26	.80	.345

Panel A *Varianzanalytische Effekte (2 x 2 ANVOA)*

Effekt	*F*	*p*	ηp²	df
CS-Präsenz	.612	.435	.003	(1, 183)
Teilnehmendengruppen	4.203	.042	.022	(1, 183)
Interaktion	3.442	.065	.018	(1, 183)

Anmerkung. HS = Hate Speech; CS = Counter Speech; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Berichtet werden Zellenmittelwerte sowie Varianzanalysen einer zweifaktoriellen Varianzanalyse (2 × 2 ANOVA) mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

Tabelle 13

Verantwortungsgefühl nach CS-Präsenz und Teilnehmendengruppen

Bedingung	Teilnehmendengruppen	*n*	*M*	*SD*	Gesamt n	Gesamt *M*	Gesamt *SD*	Levene *p*
HS	TG-eM	17	3.37	.98	63	3.36	1.05	.417
HS	TG-M	46	3.36	1.09	63	3.36	1.05	.417
HS + CS	TG-eM	21	3.29	.85	124	3.33	.99	.417
HS + CS	TG-M	103	3.34	1.02	124	3.33	.99	.417
Gesamt	TG-eM	38	3.32	.90	187	3.34	1.01	.417
Gesamt	TG-M	149	3.35	1.04	187	3.34	1.01	.417

Panel A *Varianzanalytische Effekte (2 × 2 ANOVA)*

Effekt	*F*	*p*	ηp²	df
CS-Präsenz	.068	.794	.000	(1, 183)
Teilnehmendengruppen	.011	.916	.000	(1, 183)
Interaktion	.039	.843	.000	(1, 183)

Anmerkung. HS = Hate Speech; CS = Counter Speech; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Berichtet werden Zellenmittelwerte sowie Varianzanalysen einer zweifaktoriellen Varianzanalyse (2 × 2 ANOVA) mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

Tabelle 14

Interventionsabsicht nach CS-Präsenz und Teilnehmendengruppen

Bedingung	Teilnehmendengruppen	*n*	*M*	*SD*	Gesamt *n*	Gesamt *M*	Gesamt *SD*	Levene *p*
HS	TG-eM	17	2.82	1.29	63	2.37	1.37	.512
HS	TG-M	46	2.20	1.38	63	2.37	1.37	.512
HS + CS	TG-eM	21	2.19	1.12	124	2.19	1.21	.512
HS + CS	TG-M	103	2.19	1.24	124	2.19	1.21	.512
Gesamt	TG-eM	38	2.47	1.22	187	2.25	1.27	.512
Gesamt	TG-M	149	2.19	1.28	187	2.25	1.27	.512

Panel A *Varianzanalytische Effekte (2 × 2 ANOVA)*

Effekt	*F*	*p*	ηp²	df
CS-Präsenz	1.823	.179	.010	(1, 183)
Teilnehmendengruppen	1.764	.186	.010	(1, 183)
Interaktion	1.806	.181	.010	(1, 183)

Anmerkung. HS = Hate Speech; CS = Counter Speech; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit. Berichtet werden Zellenmittelwerte sowie Varianzanalysen einer zweifaktoriellen Varianzanalyse (2 × 2 ANOVA) mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

6.5 Explorative Zusatzanalysen

Ergänzend zu den hypothesengeleiteten Analysen wurden explorative Zusatzanalysen durchgeführt, um zu prüfen, ob sich die zentralen Variablen durch weitere soziodemografische Faktoren und Moderatoren beeinflussen lassen. Diese Analysen dienen der Kontextualisierung und Mustererkennung und werden daher ausdrücklich explorativ interpretiert.

6.5.1 Wahrgenommene soziale Identitätsbedrohung

Über alle Modellspezifikationen hinweg zeigte sich konsistent, dass die Teilnehmendengruppe der ethnischen Minderheiten mit höherer wahrgenommener sozialer Identitätsbedrohung verbunden ist. Dieser Effekt erwies sich als robust und blieb auch unter Kontrolle unterschiedlicher Moderatoren (darunter Voreinstellungen gegenüber ethnischen Minderheiten, politische Orientierung sowie Social-Media-Nutzungsmerkmale) bestehen. Die Effektstärken der Teilnehmendengruppen lagen dabei durchgängig im moderaten Bereich.

Demgegenüber zeigten sich Effekte der CS-Präsenz (HS vs. HS + CS) sowie der Interaktion mit den Teilnehmendengruppen weniger stabil. Zwar traten diese Effekte in mehreren Modellspezifikationen auf, ihre Effektstärken fielen aber durchweg klein aus und waren nicht in allen Modellen signifikant. Entsprechend werden diese Befunde als vorsichtige explorative Hinweise interpretiert.

Zusätzlich zeigten sich in einzelnen Modellen Effekte des Geschlechts sowie der Kommentierhäufigkeit in sozialen Medien. Diese Effekte variierten zwischen den

Spezifikationen und werden daher nicht als robust bewertet. Die Annahmen der Varianzanalysen waren in allen Modellen unauffällig. Die vollständigen Modellübersichten sind im Anhang (siehe Tabelle D2) dokumentiert.

6.5.2 Stereotypbezogene Zuschreibungen (Eigenschaftszuschreibung und Distanz/Vermeidung)

Für die stereotypbezogenen Zuschreibungen wurden multivariate GLM-Analysen mit Eigenschaftszuschreibung und Distanz/Vermeidung berechnet. Insgesamt ergab sich kein konsistentes Muster für einen Haupteffekt der Experimentalbedingung (HS vs. HS + CS) auf diese Variablen.

Über die verschiedenen Modellspezifikationen hinweg zeigten sich für die Eigenschaftszuschreibung vor allem Alter sowie die jeweiligen Einstellungs- und Nutzungsvariablen als erklärungsrelevant. Ältere Teilnehmende sowie Personen mit positiveren Voreinstellungen gegenüber ethnischen Minderheiten bzw. einer stärker linksorientierten politischen Selbstverortung berichteten tendenziell positivere Eigenschaftszuschreibungen. Die Effektstärken dieser Prädiktoren lagen im kleinen bis moderaten Bereich. Effekte der CS-Präsenz oder der Teilnehmendengruppen traten hingegen nicht stabil auf. Einzelne Interaktionseffekte, etwa zwischen CS-Präsenz und Teilnehmendengruppen, wurden in einzelnen Modellen beobachtet, erwiesen sich allerdings als modellabhängig und nicht robust.

Für Distanz/Vermeidung zeigten sich stärker kontextabhängige und interaktive Muster. Insbesondere Voreinstellungen gegenüber ethnischen Minderheiten sowie politische Orientierung standen in mehreren Modellen mit Distanz-/Vermeidungswerten in Zusammenhang. Ferner traten vereinzelt Interaktionen zwischen CS-Präsenz, Teilnehmendengruppen und Einstellungs- bzw. Nutzungsvariablen auf. Diese Effekte variierten deutlich zwischen den Modellen. In den multivariaten Analysen zeigten sich zudem Hinweise auf Annahmeverletzungen (signifikanter Box-Test, vereinzelt verletzte Varianzhomogenität), wodurch diese Interaktionseffekte ausschließlich explorativ interpretiert werden. Eine robuste Bestätigung dieser Muster im hypothesengeleiteten Modell erfolgte nicht. Die vollständigen Modellübersichten sind im Anhang (siehe Tabelle D3 und Tabelle D4) dokumentiert.

6.5.3 Interventionsbereitschaft

Auch für die Dimensionen der Interventionsbereitschaft ergaben sich in den explorativen Analysen keine stabilen Effekte der CS-Präsenz.

Für die Schweregradwahrnehmung zeigte sich ein Zusammenhang mit den Teilnehmendengruppen sowie mit dem Alter der Teilnehmenden. Teilnehmende der Mehrheitsgesellschaft und ältere Personen bewerteten den Vorfall tendenziell als schwerwiegender. Die Effektstärken lagen hierbei im kleinen bis moderaten Bereich. Effekte der CS-Präsenz oder stabile Interaktionen traten nicht auf.

Für Verantwortungsgefühl sowie für die Interventionsabsicht ergaben sich weder im Gesamtmodell noch für einzelne Prädiktoren signifikante Effekte. Auch hier zeigten sich keine konsistenten Hinweise darauf, dass Counter Speech unter Berücksichtigung zusätzlicher Personenmerkmale systematisch mit handlungsbezogenen Reaktionen verbunden ist. Die Modellannahmen waren unauffällig. Die vollständigen Übersichten sind im Anhang (siehe Tabellen D5, D6 und D7) dokumentiert.

6.6 Zusammenfassung der zentralen Ergebnisse

Im Folgenden werden die zentralen Ergebnisse der Hypothesentests und der explorativen Forschungsfrage in komprimierter Form zusammengefasst. Eine tabellarische Übersicht ist Tabelle 15 zu entnehmen.

Für die wahrgenommene soziale Identitätsbedrohung (H1) zeigte sich, dass Counter Speech insgesamt mit geringerer wahrgenommener Bedrohung einherging. Dieser Effekt wurde durch die Teilnehmendengruppe moderiert und trat ausschließlich bei Teilnehmenden ethnischer Minderheiten auf. Damit wird H1 unterstützt.

Für die CS-Quelle im Hinblick auf wahrgenommene soziale Identitätsbedrohung (H1a) ergaben sich keine systematischen Unterschiede. Die Quelle der Counter Speech und deren Interaktion mit der Teilnehmendengruppe beeinflussten die wahrgenommene Bedrohung nicht. H1a wird daher nicht unterstützt.

Für stereotypbezogene Zuschreibungen (H2) zeigten sich weder multivariat noch in den univariaten Analysen signifikante Effekte der CS-Präsenz oder der Teilnehmendengruppe. Counter Speech veränderte stereotype Zuschreibungen somit nicht systematisch. H2 wird nicht unterstützt.

Auch im Hinblick auf die CS-Quelle (H2a) ergaben sich keine signifikanten Effekte auf stereotype Zuschreibungen im hypothesengeleiteten Modell. Einzelne Befunde erwiesen sich als nicht robust. H2a wird daher nicht unterstützt.

Für die explorative Forschungsfrage zur Interventionsbereitschaft (EF1) zeigten sich keine Hinweise darauf, dass Counter Speech Verantwortungszuschreibung oder Interventionsabsicht beeinflusst. Lediglich für die Schweregradwahrnehmung zeigte sich ein Unterschied zwischen den Teilnehmendengruppen, wobei Teilnehmende der

Mehrheitsgesellschaft den Vorfall als schwerwiegender bewerteten als Teilnehmende ethnischer Minderheiten. Insgesamt ergaben sich somit keine systematischen Zusammenhänge zwischen CS-Präsenz und handlungsbezogenen Reaktionen.

Tabelle 15

Übersicht der Hypothesentests und der explorativen Forschungsfrage

Hypothese / Forschungsfrage	Abhängige Variable(n)	Zentrale Prädiktoren	Kernergebnis (inhaltlich)	Unterstützung
H1 Wahrgenommene soziale Identitätsbedrohung	Wahrgenommene soziale Identitätsbedrohung	CS-Präsenz (HS vs. HS + CS), Teilnehmendengruppen	Counter Speech war mit geringerer wahrgenommenen sozialen Identitätsbedrohung bei Teilnehmenden ethnischer Minderheiten assoziiert	Unterstützt
H1a CS-Quelle	Wahrgenommene soziale Identitätsbedrohung	CS-Quelle, Teilnehmendengruppen	Keine Effekte der CS-Quelle	Nicht unterstützt
H2 Stereotypbezogene Zuschreibungen	Eigenschaftszuschreibung, Distanz/Vermeidung	CS-Präsenz, Teilnehmendengruppen	Keine robusten Effekte	Nicht unterstützt
H2a CS-Quelle	Eigenschaftszuschreibung, Distanz/Vermeidung	CS-Quelle, Teilnehmendengruppen	Keine robusten Effekte	Nicht unterstützt
EF1 Interventionsbereitschaft (explorativ)	Schweregradwahrnehmung, Verantwortungsgefühl, Interventionsabsicht	CS-Präsenz, Teilnehmendengruppen	Kein CS-Effekt; Teilnehmendengruppenunterschied in Schweregradwahrnehmung	Explorativ, kein CS-Effekt

Anmerkung. HS = Hate Speech; CS = Counter Speech; CS-M = Counter Speech aus der Mehrheitsgesellschaft; CS-eM = Counter Speech aus der ethnischen Minderheit; TG-M = Mehrheitsgesellschaft; TG-eM = ethnische Minderheit.

7 Diskussion

Dieses Kapitel diskutiert die Ergebnisse der vorliegenden Studie im Hinblick auf die in Kapitel 4 formulierte Forschungsfrage sowie die daraus abgeleiteten Hypothesen und die explorative Forschungsfrage. Ziel ist es, die in Kapitel 6 berichteten Ergebnisse theoriegeleitet einzuordnen, ihre Bedeutung im Kontext des bisherigen Forschungsstands zu bewerten und daraus Limitationen sowie Implikationen abzuleiten.

7.1 Einordnung der zentralen Ergebnisse in die relevante Theorie

Die Ergebnisse zeigen ein differenziertes Wirkprofil über die drei untersuchten Wirkungsebenen hinweg (affektiv-gruppenbezogen, kognitiv, handlungsbezogen). Während sich auf der affektiv-gruppenbezogenen Ebene Effekte zeigen, bleiben entsprechende Wirkungen auf der kognitiven Ebene stereotypbezogener Zuschreibungen sowie auf der handlungsbezogenen Ebene der Interventionsbereitschaft aus. Dieses Muster stützt die Annahme, dass diese Ebenen unterschiedlichen psychologischen Mechanismen folgen und nicht notwendigerweise parallel auf dieselben kommunikativen Stimuli reagieren.

Wahrgenommene soziale Identitätsbedrohung (Affektiv-gruppenbezogene Wirkungsebene). Der zentrale Befund der Studie betrifft die wahrgenommene soziale Identitätsbedrohung. Die Präsenz von Counter Speech ging im Experiment mit einer geringeren wahrgenommenen sozialen Identitätsbedrohung einher, jedoch ausschließlich bei Teilnehmenden ethnischer Minderheiten. Damit zeigt sich ein gruppenspezifischer Effekt der CS-Präsenz.

Dieser Befund ist anschlussfähig an Ansätze sozialer Identitätsbedrohung. Hate Speech kann in diesem Kontext als identitätsrelevante Abwertung verarbeitet werden, wenn die adressierte Gruppe Teil der eigenen sozialen Identität ist. Die Ergebnisse sprechen somit dafür, dass Hate Speech nicht nur als Normverletzung wahrgenommen wird, sondern für Betroffene bzw. Gruppenangehörige als identitätsbezogene Bedrohung erlebt werden kann.

Dass sich kein entsprechender Effekt bei Teilnehmenden der Mehrheitsgesellschaft zeigt, ist theoretisch plausibel. Für diese Gruppe ist die im Stimulus dargestellte Abwertung nicht in gleichem Maße identitätsrelevant. Hate Speech wird hier eher als allgemeine Normverletzung wahrgenommen, nicht hingegen als direkte Bedrohung der eigenen sozialen Identität.

Die Ergebnisse sind damit vereinbar, dass Counter Speech in diesem Kontext als sichtbares normatives Unterstützungssignal interpretiert wird. Durch die öffentliche Zurückweisung wird die kommunikative Situation so gerahmt, dass die Hate Speech weniger

als sozial akzeptierte Position erscheint. Diese veränderte Einordnung geht bei Teilnehmenden ethnischer Minderheiten mit geringerer wahrgenommener sozialer Identitätsbedrohung einher. Gleichzeitig zeigt sich kein paralleler Effekt auf kognitiver oder handlungsbezogener Ebene.

Die Reduktion wahrgenommener sozialer Identitätsbedrohung ist dabei als kurzfristiger, situationsbezogener Effekt im unmittelbaren Rezeptionskontext zu verstehen. Aussagen darüber, ob Counter Speech langfristige Diskriminierungserfahrungen oder wiederholte Exposition gegenüber Hate Speech kompensieren kann, lassen sich daraus nicht ableiten.

Hypothese H1a, die Unterschiede in Abhängigkeit von der CS-Quelle erwartete, wird hingegen nicht bestätigt. Der Befund spricht dafür, dass im gewählten Kommentarsetting weniger die Positionierung der CS-Quelle entscheidend ist als die sichtbare Funktion der Counter Speech als Gegensignal. Für die affektiv-gruppenbezogene Verarbeitung scheint somit primär relevant zu sein, dass widersprochen wird, nicht von wem.

Stereotypbezogene Zuschreibungen (Kognitive Wirkungsebene). Im Gegensatz zur affektiv-gruppenbezogenen Wirkungsebene lässt sich für die kognitive Ebene stereotypbezogener Zuschreibungen keine Hinweise auf Effekte der CS-Präsenz feststellen. Auch Unterschiede in Abhängigkeit von der CS-Quelle bleiben aus.

Dieses Ergebnis lässt sich angesichts der Priming-Ansätze einordnen. Stereotypbezogene Zuschreibungen sind häufig stärker dispositionell verankert als situative affektiv-gruppenbezogene Reaktionen. Eine einmalige, kurze Exposition gegenüber einem einzelnen Kommentar stellt daher einen vergleichsweise schwachen Stimulus für messbare Veränderungen expliziter Zuschreibungen dar. Alternativ könnten die Nullbefunde auch auf begrenzte statistische Power in TG-eM, auf Deckeneffekte positiver Einstellungen oder auf eine zu geringe Differenzierung der Counter Speech zurückgehen.

Hinzu kommt, dass die verwendeten Maße auf Selbstauskünften beruhen und bei sensiblen Themen anfällig für soziale Erwünschtheit sind. Kurzfristige kognitive Aktivierungen könnten daher unter Umständen nicht in expliziten Bewertungen sichtbar werden, auch wenn entsprechende Prozesse theoretisch möglich sind.

Diese Einordnung passt zu den explorativen Zusatzanalysen. Stereotypbezogene Zuschreibungen hängen stärker mit individuellen Voreinstellungen (z. B. politische Orientierung, Alter, vorbestehende Einstellungen gegenüber ethnischen Minderheiten) zusammen als mit der experimentellen Manipulation. Kognitive Reaktionen erscheinen damit stärker dispositionsabhängig als durch die kurzfristige Stimulusvariation erklärbar.

Zudem sind die Ausgangsniveaus in der Stichprobe zu berücksichtigen. Eigenschaftszuschreibungen fallen insgesamt eher positiv aus, Distanz ist gering und zeigt eine ausgeprägte Bodenverteilung. Vor diesem Hintergrund erscheint es plausibel, dass sich unter diesen Bedingungen keine messbaren Priming-Effekte nachweisen lassen.

Interventionsbereitschaft (Handlungsbezogene Wirkungsebene). Auch auf der handlungsbezogenen Wirkungsebene zeigen sich keine Effekte der CS-Präsenz auf Interventionsbereitschaft (Schweregradwahrnehmung, Verantwortungsgefühl oder Interventionsabsicht). Diese Befunde lassen sich vor dem Hintergrund einordnen, dass Intervention in digitalen Kommentaröffentlichkeiten ein mehrstufiger und kontextabhängiger Prozess ist.

In digitalen Kommentaröffentlichkeiten kann sichtbare Counter Speech ambivalent wirken: Sie kann als Normsignal verstanden werden, zugleich aber auch den Eindruck erzeugen, dass bereits ausreichend reagiert wurde. Die ausbleibenden Effekte deuten darauf hin, dass sich diese Ambivalenz im untersuchten Setting nicht eindeutig zugunsten erhöhter Interventionsbereitschaft auswirkt.

Auch hier liefert die Stichprobe einen relevanten Kontext. Die Mehrheit der Teilnehmenden nutzt soziale Medien primär rezeptiv und kommentiert nur selten aktiv. Entsprechend zeigen die explorativen Analysen, dass Nutzungsmuster (insbesondere Kommentierhäufigkeit) eher mit einzelnen Bewertungen zusammenhängen als die CS-Präsenz. Handlungsbezogene Reaktionen erscheinen damit stärker an Routinen und Nutzungspraxen gebunden als durch die kurzfristige Präsenz von Counter Speech erklärbar.

Die Befunde verdeutlichen somit, dass eine Reduktion wahrgenommener sozialer Identitätsbedrohung durch Counter Speech nicht automatisch in erhöhte Interventionsbereitschaft übersetzt wird. Affektiv-gruppenbezogene Entlastung und handlungsbezogene Aktivierung folgen offenbar unterschiedlichen Prozessen.

Beantwortung der Forschungsfrage. In Bezug auf die übergeordnete Forschungsfrage zeigt sich ein wirkungsebenenspezifisches Muster. Die Counter Speech reduziert die wahrgenommene soziale Identitätsbedrohung signifikant, allerdings ausschließlich bei Teilnehmenden ethnischer Minderheiten. Für stereotypbezogene Zuschreibungen sowie für Interventionsbereitschaft lassen sich hingegen keine systematischen Effekte der CS-Präsenz nachweisen. Unterschiede in Abhängigkeit von der CS-Quelle ergaben sich ebenfalls nicht.

Insgesamt deutet die Studie darauf hin, dass Counter Speech unter den untersuchten Bedingungen vor allem als affektiv wirksames normatives Unterstützungssignal

wahrgenommen wird. Kognitive und handlungsbezogene Reaktionen erweisen sich demgegenüber als stärker dispositions- und kontextabhängig und weniger durch eine einmalige Exposition gegenüber Hate Speech und Counter Speech beeinflussbar.

7.2 Einordnung im Kontext des bisherigen Forschungsstands und der Forschungslücke

Vor dem Hintergrund dieser Einordnung erfolgt die Einbettung der Ergebnisse in den bisherigen Forschungsstand sowie in die in Kapitel 3 identifizierten Forschungslücken.

Frühere Studien weisen wiederholt auf ein heterogenes und kontextabhängiges Wirkprofil von Counter Speech hin. Die vorliegende Studie trägt zur Präzisierung dieser Befundlage bei, indem sie Hate Speech und Counter Speech innerhalb eines realitätsnahen Kommentarsettings experimentell vergleicht, mehrere Wirkungsebenen gleichzeitig erfasst und die Teilnehmendengruppen systematisch berücksichtigt. Damit rückt stärker in den Fokus, unter welchen Bedingungen und auf welcher Wirkungsebene sich Effekte zeigen.

Die Ergebnisse ermöglichen zudem eine differenziertere Einordnung der in Kapitel 3 identifizierten Forschungslücken. Insbesondere verdeutlichen sie, dass Abweichungen in der berichteten Wirksamkeit von Counter Speech auch daraus zurückzuführen sein können, dass verschiedene Studien unterschiedliche Wirkungsebenen betrachten. Während frühere Arbeiten affektiv-gruppenbezogene, kognitive oder handlungsbezogene Effekte oft isoliert untersuchten, zeigt die gleichzeitige Betrachtung in dieser Studie, dass diese Prozesse empirisch nicht parallel verlaufen müssen.

Ferner unterstreichen die gruppenspezifischen Effekte auf wahrgenommene soziale Identitätsbedrohung die Bedeutung sozialer Gruppenzugehörigkeit. Studien, die soziale Gruppen nicht getrennt betrachten, können zentrale Wirkmechanismen von Hate Speech und Counter Speech übersehen. Die Ergebnisse sprechen dafür, affektiv-gruppenbezogene Wirkungen systematisch in Abhängigkeit davon zu untersuchen, ob Personen selbst der abgewerteten Gruppe angehören.

Schließlich erlaubt der experimentelle Vergleich von Hate Speech mit und ohne Counter Speech innerhalb desselben Kommentarsettings eine differenziertere Einordnung ambivalenter Befunde zur Wirksamkeit von Counter Speech. Die Ergebnisse sprechen eher für eine wirkungsebenenspezifische Betrachtung als für ein allgemeines Wirkmodell, in dem Counter Speech entweder generell wirksam oder generell wirkungslos ist.

7.3 Methodische und konzeptionelle Limitationen der Studie

Die Ergebnisse der vorliegenden Studie sind vor dem Hintergrund mehrerer methodischer und konzeptioneller Limitationen zu interpretieren. Diese betreffen insbesondere die Ausgestaltung des experimentellen Settings, die Stichprobenstruktur, die Operationalisierung

zentraler Konstrukte sowie das analytische Vorgehen. Die folgenden Punkte dienen der Einordnung, welche Wirkprozesse mit dem gewählten Design empirisch erfasst werden konnten und welche darüber hinausgehen.

Eine zentrale Limitation ergibt sich aus der Gestaltung des experimentellen Stimulusmaterials. Die Studie basiert auf einer einmaligen, kurzen Exposition gegenüber einem einzelnen Hate-Speech-Kommentar in einem Instagramähnlichen Kommentarsetting. Dieses Vorgehen eignet sich, kurzfristige affektiv-gruppenbezogene Reaktionen unter kontrollierten Bedingungen zu erfassen. Es bildet allerdings nicht die Dynamik wiederholter Expositionen oder eskalierender Kommentarverläufe ab, wie sie in realen digitalen Kommunikationsumgebungen häufig vorkommen. Gerade kognitive und handlungsbezogene Wirkungen dürften stärker von Wiederholung, Intensität und situativer Einbettung abhängen, als es ein einzelner Stimulus abbilden kann. Entsprechend sind ausbleibende Effekte auf stereotypbezogene Zuschreibungen und Interventionsbereitschaft vor diesem Hintergrund vorsichtig zu interpretieren.

Zudem wurde im Anschluss an die Stimuluspräsentation kein expliziter Manipulationscheck durchgeführt, der systematisch erfasst hätte, ob die dargestellten Kommentare eindeutig als Hate Speech oder Counter Speech erkannt und eingeordnet wurden. Auch wenn einzelne Befunde darauf hindeuten, dass die Stimuli überwiegend im erwarteten Sinne verstanden wurden, bleibt offen, ob alle Teilnehmenden die Inhalte in gleicher Weise interpretiert haben.

Eine weitere Limitation betrifft die Stichprobenstruktur. Die Rekrutierung über soziale Medien und persönliche Netzwerke führte zu einer freiwilligen Online-Stichprobe mit typischen Merkmalen von Selbstselektion. Die Teilnehmenden weisen insgesamt eine eher minderheitenfreundliche Grundhaltung sowie ein hohes Bildungsniveau auf. Diese Voraussetzungen können zu eingeschränkter Varianz in zentralen Einstellungsmaßen führen und insbesondere kognitive Effekte abschwächen, wenn Ausgangsniveaus bereits relativ positiv oder homogen sind.

Konzeptionelle Einschränkungen ergeben sich zudem aus der Operationalisierung zentraler Begriffe. Der verwendete Begriff ethnische Minderheiten schließt unterschiedliche Gruppen mit potenziell heterogenen Erfahrungen ein, wodurch die in Kapitel 2.2 betonte macht- und positionsbezogene Definition im Fragebogen nur eingeschränkt trennscharf abgebildet werden konnte. Es ist nicht auszuschließen, dass dieser Begriff von Teilnehmenden unterschiedlich interpretiert wurde, was die inhaltliche Präzision der Messungen begrenzen kann.

Auch die Erhebungslogik der abhängigen Variablen bringt Einschränkungen mit sich. Die Messung erfolgte über Selbstauskünfte unmittelbar nach der Stimuluspräsentation. Dieses Vorgehen ist für subjektive Wahrnehmungen angemessen, erlaubt jedoch keine Aussagen über tatsächliches Verhalten in realen Kommunikationssituationen. Insbesondere handlungsbezogene Reaktionen wurden als Intentionen und Verantwortungszuschreibungen erfasst, nicht hingegen als beobachtbares Interventionsverhalten. Zudem können kurzfristige kognitive Aktivierungen nicht zwingend in expliziten Bewertungen sichtbar werden.

Auch das gewählte Auswertungsverfahren bringt Einschränkungen mit sich. Die eingesetzten varianzanalytischen Verfahren sind vor allem darauf ausgelegt, durchschnittliche Unterschiede zwischen Gruppen zu untersuchen. Komplexere Wirkzusammenhänge, etwa indirekte Effekte oder nichtlineare Verläufe, lassen sich damit nur begrenzt erfassen. Kleinere Effekte könnten unter diesen Bedingungen unentdeckt geblieben sein.

Für den Kontrast HS vs. HS + CS zeigte sich ein signifikanter Unterschied in der Geschlechtsverteilung, weshalb es vorsichtig interpretiert werden muss. Aufgrund kleiner erwarteter Zellhäufigkeiten ist dieser Befund vorsichtig zu interpretieren; zugleich stellt er eine potenzielle Einschränkung der Randomisierung dar und sollte bei der Einordnung der Effekte berücksichtigt werden.

Die genannten Limitationen schmälern nicht den zentralen Beitrag der Studie, sondern ordnen dessen Aussagekraft ein. Die Ergebnisse beziehen sich auf spezifische Wirkungsebenen in klar definierten Rahmenbedingungen. Diese Einordnung bildet den Ausgangspunkt für die im folgenden Abschnitt abgeleiteten Implikationen für zukünftige Forschung.

7.4 Implikationen für zukünftige Forschung

Die Ergebnisse der vorliegenden Studie legen nahe, Wirkungen von Hate Speech und Counter Speech in digitalen Kommentaröffentlichkeiten künftig konsequenter als mehrdimensionale, kontextabhängige und gruppenspezifische Prozesse zu untersuchen. Die differenzierten Befunde über die drei Wirkungsebenen verdeutlichen, dass pauschale Aussagen zur Wirksamkeit von Counter Speech theoretisch verkürzt sind. Zukünftige Forschung sollte daher die verschiedenen Wirkungsebenen systematisch unterscheiden und ihre Zusammenhänge gezielt modellieren.

Vor dem Hintergrund der Effekte auf wahrgenommene soziale Identitätsbedrohung erscheint es besonders relevant, soziale Identität stärker als zentralen Bestandteil digitaler Wirkprozesse zu konzeptualisieren. Die Befunde deuten darauf hin, dass Hate Speech vor allem dann affektiv wirksam wird, wenn sie an für Mitlesende selbstbedeutsame

Identitätskategorien anschließt. Counter Speech scheint dabei weniger als argumentative Korrektur zu wirken, sondern primär als sichtbares soziales Unterstützungssignal. Für die Weiterentwicklung theoretischer Modelle ergibt sich daraus die Implikation, affektiv-gruppenbezogene Wirkungen stärker als situativ eingebettete soziale Identitätsprozesse zu verstehen. Zukünftige Studien könnten hier differenzieren zwischen kollektiver Identitätsbedrohung, personaler Abwertung und wahrgenommener Normverletzung. Ergänzend erscheint es sinnvoll, affektive Reaktionen nicht ausschließlich über Selbstberichte, sondern auch über implizite oder physiologische Maße zu erfassen.

Die ausbleibenden Effekte auf stereotypbezogene Zuschreibungen legen nahe, dass kurzfristige situative Reize nicht zwangsläufig zu messbaren Veränderungen expliziter Bewertungen führen. Zukünftige Forschung sollte daher kognitive Aktivierungsprozesse klarer von stabilen Einstellungsdimensionen abgrenzen und verstärkt auf indirekte oder implizite Messverfahren zurückgreifen. Ferner sprechen die Befunde dafür, wiederholte Expositionen, kumulative Erfahrungen oder längsschnittliche Designs einzubeziehen, um mögliche Normverschiebungen oder Desensibilisierungseffekte untersuchen zu können.

Auch im Bereich der Interventionsbereitschaft (Schweregradwahrnehmung, Verantwortungsgefühl, Interventionsabsicht) verdeutlichen die Ergebnisse, dass handlungsbezogene Prozesse in digitalen Öffentlichkeiten nicht unmittelbar aus affektiven oder kognitiven Reaktionen abgeleitet werden können. Zukünftige Forschung sollte Modelle des Interventionsverhaltens stärker an die spezifischen Bedingungen digitaler Kommunikationsräume anpassen. Kommentarbereiche sind durch Anonymität, unklare Publikumsgrößen und vielfältige Interventionsformen geprägt. Entsprechend scheint es sinnvoll, zwischen verschiedenen Formen digitaler Intervention (z. B. öffentliches Kommentieren, Melden von Inhalten, unterstützende Reaktionen) zu unterscheiden und Interventionsbereitschaft nicht nur als Intention, sondern auch als beobachtbares Verhalten in realitätsnahen oder feldexperimentellen Designs zu untersuchen.

Obwohl in der vorliegenden Studie keine Effekte der CS-Quelle nachweisbar waren, sollten mögliche Unterschiede zwischen verschiedenen Sprecher*innenpositionen nicht vorschnell ausgeschlossen werden. Die Befunde legen nahe, dass im untersuchten Setting die bloße Präsenz von Counter Speech als Normsignal dominanter wirkte als die Gruppenzugehörigkeit der Quelle. Zukünftige Forschung könnte analysieren, in welchen Situationen die Auswirkungen der CS-Quelle besonders deutlich werden, beispielsweise in stark polarisierten Umgebungen oder bei offensichtlichen Gruppenmerkmalen der Sprecher*innen.

Ferner erscheint es sinnvoll, individuelle Voreinstellungen, politische Orientierung und Social-Media-Nutzungstypen sozialer Medien nicht nur als Kontrollvariablen, sondern als theoretisch relevante Moderatoren zu modellieren. Auf diese Weise lassen sich unterschiedliche Wirkverläufe genauer erfassen, insbesondere im Zusammenspiel zwischen situativen Reizen und personengebundenen Merkmalen.

Schließlich unterstreichen die Befunde den Bedarf an methodischer Weiterentwicklung. Zukünftige Studien sollten verstärkt mehrfache oder längere Expositionen, realistischere Kommunikationsverläufe sowie prozessuale Modellierungen einsetzen, um komplexe Wirkzusammenhänge differenzierter abzubilden. Ebenso erscheint die Integration von Manipulationschecks zur Wahrnehmung und normativen Einordnung von Counter Speech sinnvoll, um Interpretationsspielräume zu verringern.

Insgesamt legen die Ergebnisse nahe, Hate Speech und Counter Speech nicht als isolierte Stimuli zu untersuchen. Stattdessen sollten sie als Bestandteile dynamischer, sozial eingebetteter Kommunikationsprozesse betrachtet werden, in denen Identität, Normwahrnehmung, individuelle Voreinstellungen und plattformspezifische Strukturen zusammenwirken.

7.5 Praktische Implikationen

Die Ergebnisse der vorliegenden Studie legen nahe, Erwartungen an Counter Speech in digitalen Kommentaröffentlichkeiten differenziert zu betrachten statt von einer einheitlichen Wirkung auszugehen. Counter Speech zeigt kein einheitliches Wirkprofil, sondern entfaltet ihre Wirkung vor allem als sichtbares soziales Signal in konkreten Kommunikationssituationen. Daraus lassen sich vorsichtige, aber dennoch praxisrelevante Schlussfolgerungen für den Umgang mit Hate Speech ableiten.

Die Befunde legen nahe, dass Counter Speech insbesondere für von Hate Speech betroffene Personen eine unterstützende Funktion erfüllen kann. Sichtbarer Widerspruch gegen gruppenbezogene Abwertung signalisiert, dass solche Äußerungen nicht unwidersprochen bleiben und nicht als sozial akzeptierte Norm gelten. Initiativen, die zu solidarischer Counter Speech ermutigen, können daher dazu beitragen, Unterstützung öffentlich sichtbar zu machen und digitale Räume als nicht durchgehend zustimmende Umgebungen erfahrbar werden zu lassen.

Gleichzeitig verdeutlichen die Ergebnisse, dass diese Wirkung primär in einer situativen Reduktion wahrgenommener sozialer Identitätsbedrohung besteht. Counter Speech erscheint weniger als Instrument zur unmittelbaren Veränderung von Einstellungen oder Verhaltensabsichten, sondern als kommunikative Form normativer Grenzziehung. Praktische

Strategien sollten Counter Speech daher nicht als Allzwecklösung gegen Hate Speech verstehen, sondern als einen Baustein innerhalb von verschiedenen Maßnahmen.

Da keine stabilen Effekte auf stereotypbezogene Zuschreibungen oder Interventionsbereitschaft nachweisbar waren, sollte Counter Speech nicht als alleinstehendes Mittel zur Vorurteilsreduktion oder zur Förderung aktiver Intervention überschätzt werden. Maßnahmen, die auf langfristige Änderungen der Einstellungen abzielen oder die Förderung von Zivilcourage zum Ziel haben, erfordern zusätzliche Ansätze wie Bildungsangebote, moderierte Austauschformate oder strukturelle Moderationsmaßnahmen durch Plattformen.

Die Ergebnisse unterstreichen zudem, dass Counter Speech strukturelle Moderationsmaßnahmen nicht ersetzen kann. Auch wenn sie unterstützend wirken kann, führt ihre bloße Präsenz nicht automatisch zu erhöhter Interventionsbereitschaft oder kollektiver Gegenwehr. Plattformbetreiber sollten Counter Speech daher als ergänzendes Element betrachten, nicht als primären Mechanismus zur Regulierung von Hate Speech.

Da sich Wirkungen je nach sozialer Gruppenzugehörigkeit unterscheiden, sprechen die Befunde gegen universelle Wirkungserwartungen. Während Counter Speech für Betroffene eine entlastende Funktion erfüllen kann, zeigte sich für andere Gruppen keine vergleichbare Wirkung. Praktische Strategien gegen Hate Speech sollten daher stärker zielgruppenspezifisch entwickelt werden, statt von einheitlichen Reaktionsmustern auf Counter Speech auszugehen.

Die fehlenden Unterschiede zwischen verschiedenen CS-Quellen deuten darauf hin, dass im untersuchten Setting vor allem die sichtbare Präsenz von Counter Speech relevant war. Für die Praxis bedeutet dies, dass solidarische Counter Speech nicht zwingend von bestimmten Sprecher*innengruppen ausgehen muss, um unterstützende Effekte zu entfalten. Dieser Befund sollte jedoch nicht unkritisch verallgemeinert werden, da in anderen Kontexten differenziertere Effekte der Quelle möglich sein könnten.

Wie solche mehrdimensionalen Ansätze praktisch umgesetzt werden können, zeigen bestehende zivilgesellschaftliche Initiativen. Organisationen wie ZARA mit der Beratungsstelle #GegenHassimNetz verbinden psychosoziale Unterstützung, rechtliche Einordnung und technische Intervention, indem sie Betroffene beim Melden von Inhalten unterstützen und Plattformen Hinweise auf problematische Inhalte liefern. Dabei wird deutlich, dass viele Inhalte zwar nicht strafrechtlich relevant sind, aber dennoch als belastend erlebt werden, was die Bedeutung niedrigschwelliger Unterstützungsangebote unterstreicht. Ergänzend zielen Programme wie Web@ngels darauf ab, Ehrenamtliche in digitaler Zivilcourage und strategischer Counter Speech zu schulen, um in Online-Diskussionen sichtbar normative Grenzen zu markieren und solidarische Unterstützung zu signalisieren.

Solche Ansätze verdeutlichen, dass Counter Speech in der Praxis häufig als strukturierte, begleitete und reflektierte Form digitalen Engagements verstanden wird und in umfassendere Unterstützungs- und Präventionsstrukturen eingebettet ist.

8 Fazit und Ausblick

Die vorliegende Studie untersuchte die Wirkungen von Hate Speech und Counter Speech in digitalen Kommentaröffentlichkeiten und griff damit die eingangs formulierte Frage auf, welche Bedeutung digitale Kommunikationsräume für gesellschaftliche Aushandlungsprozesse haben. Digitale Plattformen sind zentrale Orte öffentlicher Sichtbarkeit, in denen soziale Zugehörigkeit, Anerkennung und Abwertung kommunikativ verhandelt werden. Hate Speech und Counter Speech sind in diesem Kontext nicht nur einzelne Kommunikationsakte, sondern Bestandteile öffentlicher Normaushandlungen. Die vorliegende Studie bildet dabei einen spezifischen Ausschnitt in einem kontrollierten Kommentarsetting ab.

8.1 Zentrale Schlussfolgerungen der Studie

Die Ergebnisse der Studie zeigen, dass die Wirkungen von Hate Speech und Counter Speech nicht als einheitlicher Mechanismus beschrieben werden können. Stattdessen zeigt sich ein nach Wirkungsebenen differenziertes Muster. Counter Speech erweist sich im untersuchten Kontext primär als sichtbares soziales Signal, das gruppenbezogene Abwertung kommunikativ einordnet und deren normative Akzeptanz infrage stellt. Weitergehende Effekte auf stereotypbezogene Zuschreibungen oder auf handlungsbezogene Reaktionen ergaben keine systematischen Hinweise.

Damit verweist die Studie auf einen grundlegenden Unterschied zwischen der sozialen Bedeutung von Counter Speech und ihren weiterführenden psychologischen Effekten. Counter Speech formuliert öffentlich sichtbaren Widerspruch und macht sichtbar, dass abwertende Kommunikation nicht unwidersprochen bleibt. Diese soziale Funktion ist für digitale Öffentlichkeiten zentral, in denen normative Orientierung häufig über beobachtbare Kommunikationsmuster erfolgt. Zugleich zeigt sich, dass diese symbolische und kommunikative Funktion nicht automatisch mit Veränderungen individueller Einstellungen oder Verhaltensabsichten einhergeht.

Zusammenfassend legt die Studie nahe, Hate Speech und Counter Speech nicht als spiegelbildliche Mechanismen zu verstehen. Während Hate Speech Formen sozialer Abwertung öffentlich sichtbar macht, fungiert Counter Speech vor allem als kommunikative Grenzsetzung. Beide sind Teil desselben normativen Aushandlungsprozesses, folgen aber unterschiedlichen Wirklogiken und erfüllen unterschiedliche soziale Funktionen.

8.2 Einordnung des Erkenntnisgewinns

Der zentrale Erkenntnisgewinn der Studie liegt in der differenzierten Betrachtung digitaler Kommentaröffentlichkeiten als sozial strukturierte Kommunikationsräume. Durch die

gleichzeitige Analyse affektiv-gruppenbezogener, kognitiver und handlungsbezogener Prozesse wird deutlich, dass Reaktionen auf Hate Speech und Counter Speech nicht gleichförmig verlaufen, sondern von sozialer Positionierung und psychologischer Verarbeitungsebene abhängen.

Damit trägt die Studie zu einem vertieften Verständnis der Rolle von Counter Speech in digitalen Öffentlichkeiten bei. Counter Speech erscheint weniger als direkt wirksames Instrument zur Veränderung individueller Einstellungen oder Verhaltensabsichten, sondern primär als sichtbares Zeichen normativer Grenzziehung und sozialer Positionierung. Der Beitrag der Studie besteht daher nicht darin, eine allgemeine Wirksamkeit von Counter Speech zu belegen oder zu widerlegen, sondern ihre spezifische soziale Funktion in digitalen Kommunikationsräumen genauer zu bestimmen.

Diese Perspektive verschiebt den Blick von der Frage, ob Counter Speech wirkt, hin zur Frage, welche Form von sozialer Bedeutung ihr in digitalen Öffentlichkeiten zukommt. Digitale Kommentarbereiche werden dadurch nicht nur als Orte individueller Meinungsäußerung relevant, sondern als Räume, in denen gesellschaftliche Normen, Zugehörigkeiten und Machtverhältnisse kommunikativ reproduziert und infrage gestellt werden.

8.3 Ausblick

In digitalen Öffentlichkeiten, die für viele Nutzer*innen zu zentralen Orten gesellschaftlicher Aushandlung geworden sind, bleibt Hate Speech ein sichtbares und relevantes Phänomen. Zugleich werden die Grenzen bestehender rechtlicher Regelungen und plattformspezifischer Moderationspraktiken zunehmend sichtbar, was sowohl wissenschaftliche als auch politische und zivilgesellschaftliche Akteur*innen vor neue Aufgaben stellt.

Zivilgesellschaftliche Initiativen wie die Beratungs- und Dokumentationsarbeit von ZARA verdeutlichen, dass Hate Speech für Betroffene keine abstrakte Kategorie darstellt. Vielmehr ist es eine konkrete Alltagserfahrung, die mit Belastung, Rückzug und eingeschränkter Teilhabe einhergehen kann (ZARA, 2025). Solche Befunde unterstreichen, dass Counter Speech als digitale Zivilcourage gesellschaftlich relevante Funktionen übernimmt, indem sie gruppenbezogener Abwertung öffentlich widerspricht und Betroffenen sichtbare Unterstützung signalisiert.

Gleichzeitig zeigen aktuelle Diskussionen um Plattformregulierung und Moderationspraktiken, dass Counter Speech allein nicht ausreicht, um die Verbreitung gruppenbezogener Abwertung einzudämmen. Das Spannungsfeld zwischen Meinungsfreiheit, Schutz vor Diskriminierung und unternehmerischer Plattformverantwortung bleibt eine

zentrale gesellschaftliche Herausforderung. Fragen danach, welche Inhalte entfernt, kontextualisiert oder sichtbar bleiben, sind nicht nur juristische, sondern auch normative Entscheidungen darüber, welche Formen öffentlicher Kommunikation als legitim gelten. Die Ausgestaltung dieser Rahmenbedingungen wird damit zu einem zentralen Faktor für die zukünftige Entwicklung digitaler Öffentlichkeiten.

Parallel dazu eröffnen neue technologische Ansätze Perspektiven für die Gestaltung digitaler Kommunikationsumgebungen. Forschungen zu feedbackbasierten Interface-Elementen, die emotionale Aspekte von Online-Kommunikation sichtbar machen, weisen darauf hin, dass Plattformdesign selbst zur Förderung reflektierterer Austauschformen beitragen könnte (Su et al., 2025). Solche Ansätze verschieben den Fokus von nachgelagerter Inhaltsmoderation hin zu präventiven, kontextsensitiven Gestaltungsstrategien. Zugleich machen sie deutlich, dass technische Interventionen stets auch normative Auswirkungen mit sich bringen.

Zukünftige Forschung steht vor der Aufgabe, kommunikative Praktiken, soziale Normwahrnehmungen, emotionale Verarbeitungsprozesse und plattformspezifische Strukturen stärker ganzheitlich zu betrachten. Dabei geht es nicht nur um die Wirkung einzelner Kommentare. Es geht auch um die erheblichen emotionalen Belastungsreaktionen bei Betroffenen und Mitlesenden und um die längerfristige Frage, wie digitale Kommunikationsräume gesellschaftliche Ungleichheiten verstärken, abschwächen oder neu strukturieren.

Die Auseinandersetzung mit Hate Speech und Counter Speech berührt damit grundlegende Fragen demokratischer Öffentlichkeit, sozialer Teilhabe und kommunikativer Verantwortung im digitalen Zeitalter. Forschung in diesem Feld leistet folglich nicht nur einen Beitrag zur Medienwirkungsforschung, sondern auch zur Einordnung des gegenwärtigen Strukturwandels öffentlicher Kommunikation. Vor dem Hintergrund einer digitalen Kommunikation, in der Sprache zunehmend polarisiert und gesellschaftliche Konfliktlinien sichtbar verdichtet werden, bleibt die Analyse von Hate Speech und Counter Speech zentral. Diese Analyse ist wichtig für das Verständnis, wie sich normative Grenzen, Zugehörigkeiten und Formen öffentlicher Auseinandersetzung in digitalen Öffentlichkeiten künftig entwickeln.

9 Referenzen

Álvarez-Benjumea, A. & Winter, F. (2018). Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review*, *34(3),* 223–237. https://doi.org/10.1093/esr/jcy005

Barth, N., Wagner, E., Raab, P. & Wiegärtner, B. (2023). Contextures of hate: Towards a systems theory of hate communication on social media platforms. *The Communication Review*, *26(3),* 209–252. https://doi.org/10.1080/10714421.2023.2208513

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-Speech: Examining content that challenges extremism online. *Demos*. http://www.demos. co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Bliuc, A., Faulkner, N., Jakubowicz, A. & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior, 87*, 75–86. https://doi.org/10.1016/j.chb.2018.05.026

Breakwell, G. M. & Jaspal, R. (2021). Coming Out, Distress and Identity Threat in Gay Men in the UK. *Sexuality Research And Social Policy*, *19(3),* 1166–1177. https://doi.org/10.1007/s13178-021-00608-4

Brown, A. (2017). What is Hate Speech? Part 2: Family Resemblances. *Law And Philosophy*, 36(5), 561–613. https://doi.org/10.1007/s10982-017-9300-x

Brown, R. (2000). Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology, 30(6),* 745–778. https://doi.org/10.1002/1099-0992(200011/12)30:6%253C745::AID-EJSP24%253E3.0.CO;2-O

Bührer, S., Koban, K. & Matthes, J. (2024). The WWW of digital hate perpetration: What, who, and why? A scoping review. *Computers in Human Behavior, 159*, 108321. https://doi.org/10.1016/j.chb.2024.108321

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

Chambers, J. R., Schlenker, B. R. & Collisson, B. (2013). Ideology and Prejudice. *Psychological Science*, *24(2),* 140–149. https://doi.org/10.1177/0956797612447820

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings Of The National Academy Of Sciences*, *118(9).* https://doi.org/10.1073/pnas.2023301118

Deaux, K., Bikmen, N., Gilkes, A., Ventuneac, A., Joseph, Y., Payne, Y. A. & Steele, C. M. (2007). Becoming American: Stereotype Threat effects in Afro-Caribbean Immigrant groups. *Social Psychology Quarterly*, *70(4),* 384–404. https://doi.org/10.1177/019027250707000408

Department of Communication, University of Vienna. (2019). Guidelines for ethical research at the Department of Communication. https://publizistik.univie.ac.at/fileadmin/user_upload/i_publizistik/Diverses/IRB/EG10 2019.pdf

Esau, K. (2022). Content analysis in the research field of incivility and hate speech in online communication. In *Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research: Ein Handbuch,* (S. 451–461). https://doi.org/10.1007/978-3-658-36179-2_38

Europäische Kommission. (2023). Kein Platz für Hass in Europa: EU-Kommission ruft zu Toleranz und Respekt auf. https://germany.representation.ec.europa.eu/news/kein-platz-fur-hass-europa-eu-kommission-ruft-zu-toleranz-und-respekt-auf-2023-12-06_de

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M. & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137(4),* 517–537. https://doi.org/10.1037/a0023304

Garland, J., Ghazi-Zahedi, K., Young, J., Hébert-Dufresne, L. & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11(1).* https://doi.org/10.1140/epjds/s13688-021-00314-6

Gelber, K. (2019). Differentiating hate speech: a systemic discrimination approach. *Critical Review Of International Social And Political Philosophy*, *24(4),* 393–414. https://doi.org/10.1080/13698230.2019.1576006

George, D., & Mallery, P. (2019). Reliability Analysis. In *IBM SPSS Statistics 26 Step by Step* (S. 235–246). https://doi.org/10.4324/9780429056765-19

Greitemeyer, T., Fischer, P., Kastenmüller, A. & Frey, D. (2006). Civil Courage and Helping Behavior. *European Psychologist*, *11(2),* 90–98. https://doi.org/10.1027/1016-9040.11.2.90

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F. & 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).* https://doi.org/10.1073/pnas.2116310118

Hsueh, M., Yogeeswaran, K. & Malinen, S. (2015). "Leave Your Comment Below": Can Biased Online Comments Influence Our Own Prejudicial Attitudes and Behaviors? *Human Communication Research*, *41(4),* 557–576. https://doi.org/10.1111/hcre.12059

Jia, Y. & Schumann, S. (2023). Tackling Hate speech online: The effect of counter-speech on subsequent bystander behavioral intentions. *OSF Preprints*. https://doi.org/10.33767/osf.io/9jmza

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S. & Bilz, L. (2022). 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

Khaleghipour, M., Koban, K. & Matthes, J. (2025). Listen to Me! Target Perceptions of Digital Hate: A Scoping Review of Recent Research. *Trauma Violence & Abuse*, *26(5),* 1082–1096. https://doi.org/10.1177/15248380241303725

Küpper, B., Klocke, U., & Hoffmann, L.-C. (2017). Einstellungen gegenüber lesbischen, schwulen und bisexuellen Menschen in Deutschland: Ergebnisse einer bevölkerungsrepräsentativen Umfrage. *Antidiskriminierungsstelle des Bundes.*

Latané, B., & Darley, J. M. (1970). The Unresponsive Bystander: Why Doesn't He Help? *Appleton-Century Crofts.*

Leonhard, L., Rueß, C., Obermaier, M. & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication And Media, 7(4),* 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Mackie, D. M., Devos, T. & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal Of Personality And Social Psychology*, *79(4),* 602–616. https://doi.org/10.1037/0022-3514.79.4.602

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D. & Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell Systematic Reviews, 21(1),* e70018. https://doi.org/10.1002/cl2.70018

Major, B. & O'Brien, L. T. (2004). The Social Psychology of Stigma. *Annual Review Of Psychology, 56(1),* 393–421. https://doi.org/10.1146/annurev.psych.56.091103.070137

Meta. (2025, January 7). More speech and fewer mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Obermaier, M., Fawzi, N. & Koch, T. (2014). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, *18(8),* 1491–1507. https://doi.org/10.1177/1461444814563519

Obermaier, M., Schmid, U. K. & Rieger, D. (2023). Too civil to care? How online hate speech against different social groups affects bystander intervention. *European Journal Of Criminology*, *20(3),* 817–833. https://doi.org/10.1177/14773708231156328

Obermaier, M., Schmuck, D. & Saleem, M. (2021). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, *25(9),* 2339–2358. https://doi.org/10.1177/14614448211017527

Ortutay, B. (2025). Meta is rolling back hate speech rules along with fact-checking and Zuckerberg says "recent elections" are a catalyst. *Fortune*. https://fortune.com/2025/01/09/meta-rolling-back-hate-speech-rules-zuckerberg-recent-elections-catalyst/

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. https://doi.org/10.17645/mac.v9i1.3399

Paz, M. A., Montero-Díaz, J. & Moreno-Delgado, A. (2020). Hate Speech: A Systematized review. *SAGE Open*, 10(4). https://doi.org/10.1177/2158244020973022

Peter, J. (2002). Medien-Priming — Grundlagen, Befunde und Forschungstendenzen. *Publizistik, 47(1),* 21–44. https://doi.org/10.1007/s11616-002-0002-4

Power, J. G., Murphy, S. T. & Coover, G. (1996). Priming Prejudice How Stereotypes and Counter-Stereotypes Influence Attribution of Responsibility and Credibility among Ingroups and Outgroups. *Human Communication Research*, *23(1),* 36–58. https://doi.org/10.1111/j.1468-2958.1996.tb00386.x

Ruscher, J. B. (2024). Hate speech. In *Cambridge University Press*. https://doi.org/10.1017/9781009534666

Saleem, M. & Ramasubramanian, S. (2017). Muslim Americans' Responses to Social Identity Threats: Effects of Media Representations and Experiences of Discrimination. *Media Psychology*, *22(3),* 373–393. https://doi.org/10.1080/15213269.2017.1302345

Schäfer, S., Rebasso, I., Boyer, M. M. & Planitzer, A. M. (2023). Can We Counteract Hate? Effects of Online Hate Speech and Counter Speech on the Perception of Social Groups. *Communication Research*, *51(5),* 553–579. https://doi.org/10.1177/00936502231201091

Schäfer, S., Sülflow, M. & Reiners, L. (2021). Hate Speech as an Indicator for the State of the Society. *Journal Of Media Psychology Theories Methods And Applications*, *34(1),* 3–15. https://doi.org/10.1027/1864-1105/a000294

Schemer, C. (2013). Priming, Framing, Stereotype. In *Handbuch Medienwirkungsforschung,* 153–169. https://doi.org/10.1007/978-3-531-18967-3_7

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counter speech on Facebook. In *66th ICA Annual Conference,* 1–23.

Schneiders, P. (2021). Hate Speech auf Online-Plattformen. *UFITA - Archiv für Medienrecht und Medienwissenschaft*, *85(2),* 269–333. https://doi.org/10.5771/2568-9185-2021-2-269

Sellars, A. (2016). Defining hate speech. *SSRN.* https://doi.org/10.2139/ssrn.2882244

Sharghi, S., Khalatbari, S., Laird, A., Lapidus, J., Enders, F. T., Meinzen-Derr, J., Tapia, A. L. & Ciolino, J. D. (2024). Race, ethnicity, and considerations for data collection and analysis in research studies. *Journal Of Clinical And Translational Science*, *8(1),* e182. https://doi.org/10.1017/cts.2024.632

Sökefeld, M. (2007). Problematische Begriffe: ‚Ethnizität', ‚Rasse', ‚Kultur', ‚Minderheit'. *Ethnizität und Migration. Einführung in Wissenschaft und Arbeitsfelder. Reimer Kulturwissenschaften*, 31–50.

Soral, W., Bilewicz, M. & Winiewski, M. (2017). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44(2),* 1–11. https://doi.org/10.1002/ab.21737

Sponholz, L. (2017). *Hate speech in den Massenmedien*. https://doi.org/10.1007/978-3-658-15077-8

Statistisches Bundesamt (2025). *Ein Drittel der Internetnutzenden stößt im Netz auf Hatespeech* [Pressemitteilung]. https://www.destatis.de/DE/Presse/Pressemitteilungen/2025/11/PD25_421_63.html

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52(6),* 613–629. https://doi.org/10.1037/0003-066x.52.6.613

Stephan, W. G., Ybarra, O., Martnez, C. M., Schwarzwald, J. & Tur-Kaspa, M. (1998). Prejudice toward Immigrants to Spain and Israel. *Journal Of Cross-Cultural Psychology*, *29(4),* 559–576. https://doi.org/10.1177/0022022198294004

Su, X., Zierau, N., Kim, S., Wang, A. Y. & Wambsganss, T. (2025). Emotionally Aware Moderation: The Potential of Emotion Monitoring in Shaping Healthier Social Media Conversations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2507.21089

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations, (S.* 33–37).

Van Houtven, E., Acquah, S. B., Obermaier, M., Saleem, M. & Schmuck, D. (2024). 'You Got My Back?' Severity and Counter-Speech in Online Hate Speech Toward Minority Groups. *Media Psychology*, *27(6),* 923–954. https://doi.org/10.1080/15213269.2023.2298684

Wang, C., Platow, M. J., Bar-Tal, D., Augoustinos, M., Van Rooy, D. & Spears, R. (2021). When are intergroup attitudes judged as free speech and when as prejudice? A social identity analysis of attitudes towards immigrants. *International Journal Of Psychology*, *57(4),* 456–465. https://doi.org/10.1002/ijop.12775

Wang, J. (2018). *The effects of dimensions of social media quality on user satisfaction: A cross-cultural comparison of Thai and Chinese users. In Global Journal of Arts, Humanities and Social Sciences, 6(7),* 14–44.

Waqas, A., Salminen, J., Jung, S., Almerekhi, H. & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS ONE*, *14*(9), e0222194. https://doi.org/10.1371/journal.pone.0222194

You, L. & Lee, Y. (2019). The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. *Telematics And Informatics*, 45, 101284. https://doi.org/10.1016/j.tele.2019.101284

ZARA – Zivilcourage und Anti-Rassismus-Arbeit. (2025). GegenHassImNetz Bericht 2025. https://assets.zara.or.at/media/ghinbericht/251125_ZARA_GHiN_Bericht_2025_final.pdf

Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8(2),* 127–135. https://www.surveymethods.org

Anhang

Anhang A

Abstract

Hate Speech in sozialen Medien stellt eine gesellschaftlich relevante Herausforderung dar, da sie emotionale Belastungen bei Betroffenen auslösen, Ungleichheit reproduzieren und Wahrnehmungen sozialer Normen in digitalen Öffentlichkeiten verschieben kann. Neben individuellen Folgen berührt Hate Speech damit Fragen gesellschaftlicher Teilhabe, demokratischer Öffentlichkeit und digitaler Diskussionskultur. Als öffentlich sichtbare Form digitaler Zivilcourage gewinnt Counter Speech zunehmend an Bedeutung, da sie Hate Speech widerspricht, normative Grenzen markiert und Unterstützung für Betroffene signalisiert. Empirische Befunde zu ihren Wirkungen sind jedoch bislang uneinheitlich, insbesondere in Bezug auf unterschiedliche Wirkungsebenen und soziale Gruppenzugehörigkeiten.

Die vorliegende experimentelle Online-Studie (N = 191) untersuchte in einem realitätsnahen Instagram-Kommentarsetting, wie die Präsenz von Counter Speech die wahrgenommene soziale Identitätsbedrohung, stereotypbezogene Zuschreibungen sowie die Interventionsbereitschaft beeinflusst. Zudem wurde geprüft, ob sich Effekte in Abhängigkeit von der Teilnehmendengruppe (Mehrheitsgesellschaft vs. ethnische Minderheiten) sowie der Quelle der Counter Speech unterscheiden. Die Ergebnisse zeigen ein wirkungsebenenspezifisches Muster. Counter Speech ging mit geringerer wahrgenommener sozialer Identitätsbedrohung einher, ausschließlich bei Teilnehmenden ethnischer Minderheiten. Für stereotypbezogene Zuschreibungen und Interventionsbereitschaft zeigten sich keine signifikanten Effekte. Die Quelle der Counter Speech spielte ebenfalls keine systematische Rolle. Insgesamt deuten die Befunde darauf hin, dass die Wirkung von Counter Speech kontext- und gruppenabhängig ist und nicht auf allen untersuchten Ebenen gleichermaßen greift. Die Studie unterstreicht damit die gesellschaftliche Bedeutung von Counter Speech als sichtbare Form digitaler Zivilcourage, macht zugleich deutlich, dass nachhaltige Strategien im Umgang mit Hate Speech ein Zusammenspiel aus zivilgesellschaftlichen, politischen, rechtlichen und plattformspezifischen Maßnahmen erfordern.

Abstract Englisch

Hate speech on social media poses a socially relevant challenge, as it can cause emotional distress to those affected, reproduce inequality, and shift perceptions of social norms in digital public spheres. In addition to individual consequences, hate speech thus touches on issues of social participation, democratic public discourse, and digital discussion culture. As a publicly visible form of digital civil courage, counter speech is becoming increasingly important because it contradicts hate speech, marks normative boundaries, and signals support for those affected. However, empirical findings on its effects have been inconsistent previous, especially with regard to different levels of impact and social group affiliations. This experimental online study (N = 191) examined how the presence of counter speech influences perceived social identity threat, stereotype-related attributions, and willingness to intervene in a realistic Instagram comment setting. In addition, the study examined whether effects differed depending on the participant group (majority society vs. ethnic minorities) and the source of counter speech. The results show an effect-level-specific pattern: counter speech was associated with lower perceived social identity threat, but only among participants belonging to ethnic minorities. No significant effects were found for stereotype-related attributions and willingness to intervene. The source of counter speech did not play a systematic role either. Overall, the findings suggest that the effect of counter speech is context- and group-dependent and does not work consistently at all levels examined. The study thus emphasizes the social significance of counter speech as a visible form of digital civil courage, also showing that sustainable strategies for dealing with hate speech require a combination of civil society, political, legal, and platform-specific measures.

Anhang B

Hilfsmittel

Tabelle B 1

Verwendete digitale und KI-gestützte Hilfsmittel

Hilfsmittel	Einsatzform	Beispielhafte Prompts / Nutzung
Connected Papers	Unterstützung bei der Literaturrecherche durch Visualisierung thematisch verwandter wissenschaftlicher Arbeiten	Exploration verwandter Forschungsstränge und Publikationen
DeepL	Übersetzung einzelner Fachbegriffe sowie kurzer Textpassagen zur sprachlichen Orientierung	Übersetzung deutsch–englischer Fachtermini
Zotero	Literaturverwaltung, Organisation wissenschaftlicher Quellen, automatische Erstellung von Literaturverweisen und Literaturverzeichnis im vorgegebenen Zitierstil	Generierung von Zitaten im Text, Formatierung der In-Text-Belege und des Literaturverzeichnisses
SoSci Survey	Erstellung und Durchführung einer Online-Befragung	Gestaltung des Fragebogens, Erhebung und Datenexport
IBM SPSS Statistics	Statistische Auswertung erhobener Daten, Durchführung von Analysen, Erstellung von Diagrammen und Tabellen auf Basis des eigenen Datensatzes	Anwendung statistischer Tests, Berechnung von Kennwerten, grafische Darstellung von Ergebnissen
ChatGPT-4 (Open AI)	Unterstützung bei der Strukturierung von Kapiteln, sprachliche Überarbeitung einzelner Abschnitte, Erstellung von Tabellen- oder Gliederungsentwürfen, Zusammenfassung von Texten, Unterstützung bei der Syntax-Erstellung	„Gib mir Überarbeitungsvorschläge für diesen Absatz“, „Schlage eine sinnvolle Gliederung für ein Kapitel zum Thema … vor“, „Fasse folgenden Text in Stichpunkten zusammen“, „Wie lautet die SPSS-Syntax für …?“

Anhang C

Fragebogen der Studie

Abbildung C1

Briefing

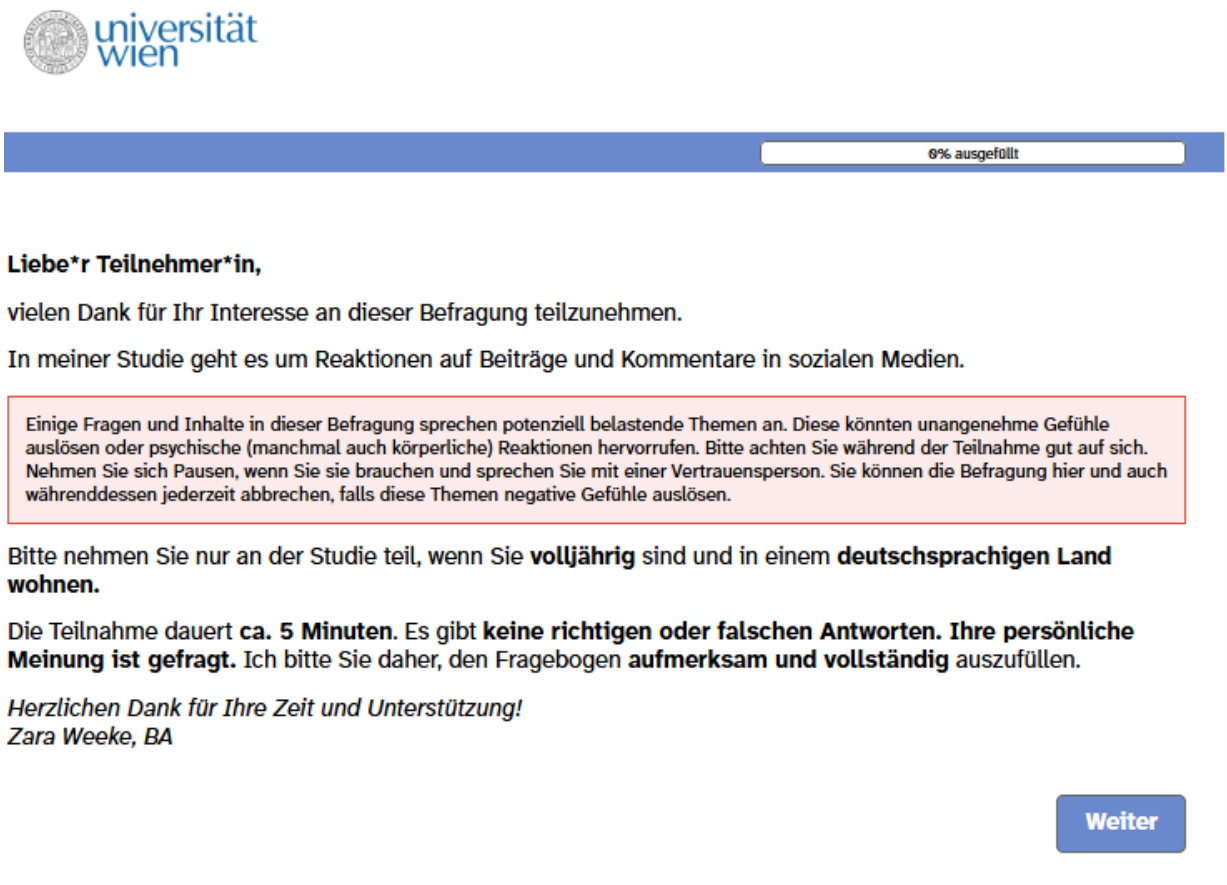


Abbildung C2

Datenschutzerklärung und Einwilligung

Datenschutzerklärung

Bevor der Fragebogen startet, sehen Sie detaillierte Informationen zu Ihren Rechten im Zuge dieser Befragung und werden nochmals um Ihre Zustimmung gebeten.

Die Daten können von der Betreuerin der wissenschaftlichen Arbeit für Zwecke der Leistungsbeurteilung eingesehen werden. Die erhobenen Daten dürfen gemäß Art. 89 Abs. 1 DSGVO grundsätzlich unbeschränkt gespeichert werden. Es besteht das Recht auf Auskunft durch die Verantwortliche an dieser Studie über die erhobenen personenbezogenen Daten sowie das Recht auf Berichtigung, Löschung, Einschränkung der Verarbeitung der Daten sowie ein Widerspruchsrecht gegen die Verarbeitung sowie des Rechts auf Datenübertragbarkeit. Ihre Daten werden ausschließlich auf Grundlage der gesetzlichen Bestimmungen (§ 2f Abs. 5 FOG) erhoben und verarbeitet. Sie verfügen über folgende persönliche Rechte im Rahmen dieser Befragung:

- Die Teilnahme an der Studie ist **freiwillig**. Sie können den Fragebogen jederzeit abbrechen.
- Ihre Teilnahme ist **anonym**, Ihre Antworten können nicht auf Sie zurückgeführt werden. Das bedeutet ebenfalls, dass Ihr persönlicher Datensatz nach Abschluss der Befragung für uns **nicht identifizierbar** ist.
- Falls Sie nach der Studie Auskunft über Ihre Daten haben wollen oder Ihre **Teilnahme zurückziehen**, bitten wir Sie, dies im abschließenden Kommentarfeld (falls nötig gemeinsam mit einer Kontaktadresse) zu vermerken.
- Deine Daten werden **ausschließlich für wissenschaftliche Zwecke** verwendet.
- Die Forschung folgt **keinem kommerziellen Interesse**. All Ihre Daten werden **streng vertraulich** behandelt.

Wenn Sie Fragen zu dieser Erhebung haben, wende Sie sich bitte gerne an die Verantwortliche dieser Untersuchung: Zara Weeke (a12339013@unet.univie.ac.at), Studentin der Studienrichtung Publizistik und Kommunikationswissenschaft an der Universität Wien.

Für grundsätzliche juristische Fragen im Zusammenhang mit der DSGVO/FOG und studentischer Forschung wende Sie sich an den Datenschutzbeauftragen der Universität Wien, Dr. Daniel Stanonik, LL.M. (verarbeitungsverzeichnis@univie.ac.at). Zudem besteht das Recht der Beschwerde bei der Datenschutzbehörde (bspw. über dsb@dsb.gv.at).

Ich freue mich, dass Sie mich bei meiner Masterarbeit unterstützen und bedanke mich im Vorfeld für Ihre Teilnahme!

Mit Ihrer Teilnahme bestätigen Sie, dass Sie die Datenschutzerklärung sowie die Teilnehmer*inneninformationen gelesen haben, diesen zustimmen und dass Sie volljährig sind.

- ◯ Ja, ich stimme zu und bin volljährig
- ◯ Nein, ich stimme nicht zu
- ◯ Ich bin noch nicht volljährig

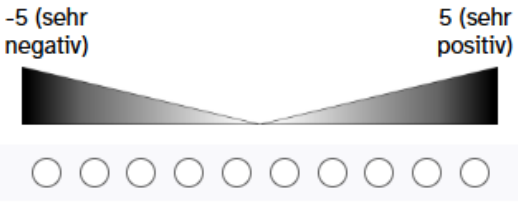
Abbildung C3

Zentrale Moderatorvariablen

Im folgenden werden ein paar Fragen gestellt. Bitte beantworten Sie diese Fragen.

Wenn Sie an ethnische Minderheiten[?] im deutschsprachigen Raum denken, wie würden Sie Ihre allgemeine Einstellung gegenüber dieser Gruppe beschreiben?

Bitte geben Sie an, wie positiv oder negativ Sie gegenüber ethnischen Minderheiten eingestellt sind. Verwenden Sie dazu eine Skala von –5 (sehr negativ) bis +5 (sehr positiv). Mit den Werten dazwischen können Sie Ihre Meinung entsprechend abstufen.



In der Politik geht es oft darum, links oder rechts zu stehen.

Wo würden Sie sich auf einer Skala von 0 (links) bis 10 (rechts) einordnen?

Wie viel Zeit verbringen Sie durchschnittlich pro Tag auf sozialen Medien?

- ◯ unter 1 Stunde
- ◯ 1 bis 2 Stunden
- ◯ 2 bis 3 Stunden
- ◯ 3 bis 4 Stunden
- ◯ 4 Stunden und mehr

Wie oft machen Sie Folgendes, wenn Sie soziale Medien nutzen?

Wählen Sie jeweils die Antwort aus, die Ihrem Verhalten am besten entspricht.

Informationen suchen	nie	selten	manchmal	oft	sehr oft
Beiträge kommentieren	nie	selten	manchmal	oft	sehr oft
Informationen teilen	nie	selten	manchmal	oft	sehr oft

Bitte schauen Sie sich den folgenden Instagram-Post an und lesen Sie **den darunter angezeigten Kommentar bzw. die Kommentare aufmerksam durch**. Im Anschluss werden Fragen dazu gestellt.

Abbildung C4

Aufmerksamkeitsfrage

In dem Instagram-Post ging es um..

- ◯ Ethnische Minderheiten
- ◯ Wahlverhalten in Deutschland

Abbildung C 5

Abhängige Variablen

Bitte lesen Sie sich folgende Fragen aufmerksam durch und beantworten Sie diese.

Nach dem Lesen eines oder mehrerer Kommentare unter dem Instagram-Post fühle ich mich...

Kreuzen Sie jeweils die Antwort an, die am besten auf Ihr Empfinden zutrifft.

	trifft überhaupt nicht zu				trifft voll zu
erniedrigt	◯	◯	◯	◯	◯
diskriminiert	◯	◯	◯	◯	◯
beleidigt	◯	◯	◯	◯	◯
bedroht	◯	◯	◯	◯	◯
in meinem Selbstwertgefühl untergraben	◯	◯	◯	◯	◯

Wenn Sie an ethnische Minderheiten[?] denken, inwieweit glauben Sie, dass die folgenden Eigenschaften auf diese Gruppe zutreffen?

Bitte geben Sie Ihre persönliche Einschätzung an.

	trifft überhaupt nicht zu				trifft voll zu
offen	◯	◯	◯	◯	◯
tolerant	◯	◯	◯	◯	◯
freundlich	◯	◯	◯	◯	◯
vertrauenswürdig	◯	◯	◯	◯	◯

Inwieweit stimmen Sie den folgenden Aussagen zu?

Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen.

	Ich stimme überhaupt nicht zu				Ich stimme voll zu
Ich möchte nichts mit ethnischen Minderheiten[?] zu tun haben.	○	○	○	○	○
Ich meide ethnische Minderheiten[?]	○	○	○	○	○
Ich halte Abstand zu ethnischen Minderheiten[?]	○	○	○	○	○

Bitte bewerten Sie, inwieweit Sie den unter dem Instagram-Post gezeigten Kommentar als bedrohlich, alarmierend oder gefährlich empfinden.

Nicht bedrohlich	○	○	○	○	○	Sehr bedrohlich
Nicht alarmierend	○	○	○	○	○	Sehr alarmierend
Nicht gefährlich	○	○	○	○	○	Sehr gefährlich

Inwieweit stimmen Sie den folgenden Aussagen zu?

Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen.

	Ich stimme überhaupt nicht zu				Ich stimme voll zu
Ich fühle mich persönlich stark verantwortlich, ethnische Minderheiten[?] zu unterstützen	○	○	○	○	○
Ich betrachte es als meine Pflicht, ethnischen Minderheiten[?] zu helfen	○	○	○	○	○
Es ist meine Verpflichtung, etwas gegen den gezeigten Kommentar zu unternehmen	○	○	○	○	○

Bitte geben Sie an, inwieweit Sie der folgenden Aussage zustimmen.

Bitte wählen Sie aus, inwieweit Sie der Aussage zustimmen.

	Ich stimme überhaupt nicht zu				Ich stimme voll zu
Ich würde auf den Kommentar reagieren	○	○	○	○	○

Abbildung C6

Soziodemografische Angaben

Zum Schluss nur noch ein paar Fragen zu Ihren demografischen Daten.

Wie alt sind Sie?

Bitte in Jahren angeben

☐ Jahre

Welchem Geschlecht fühlen Sie sich zugehörig?

[Bitte auswählen]

Welches ist der höchste Bildungsabschluss, den Sie haben?

Bitte nennen Sie Ihren höchsten Bildungsabschluss.

- ○ ohne Schulabschluss
- ○ Mittlere Reife / Sekundärschule
- ○ Abitur / Matura
- ○ Bachelor-Abschluss / Diplom / Fachhochschuldiplom
- ○ Master-Abschluss / Magister
- ○ Promotion / Doktorat / Disseration
- ○ Andere (Bitte spezifizieren)

In welchem Land leben Sie derzeit?

- ○ Deutschland
- ○ Österreich
- ○ Schweiz
- ○ Anderes, und zwar

Fühlen Sie sich einer ethnischen Minderheiten[?] im deutschsprachigen Raum zugehörig?

- ○ Ja
- ○ Nein
- ○ Keine Angabe

Haben Sie noch Anmerkungen?

Falls Sie mir noch etwas mitteilen möchten, hier können Sie es direkt hier schreiben.

Abbildung C7

Debriefing inklusive Unterstützungs- und Kontaktstelle

Liebe*r Teilnehmer*in!

Vielen herzlichen Dank für Ihre Teilnahme and meiner Befragung.

In dieser Studie wurde untersucht, wie sich **Hassrede** und verschiedene Formen von **Gegenrede** auf die Wahrnehmung sozialer Identitätsbedrohung, Stereotypenbildung und die Bereitschaft zur Intervention auswirken. Ziel ist es, besser zu verstehen, wie Online-Kommunikation das gesellschaftliche Miteinander beeinflusst und welche Maßnahmen wirkungsvoll gegen Hassrede eingesetzt werden können.

Die in der Umfrage gezeigten Beiträge (Posts) sind **nicht echt**, sondern wurden **fiktiv erstellt und ausschließlich zu Forschungszwecken für diese Studie entwickelt**. Sie spiegeln keine realen Aussagen oder Meinungen wider.

Sollten Sie sich durch die Konfrontation mit den gezeigten Inhalten emotional belastet fühlen oder Unterstützung benötigen, wenden Sie sich bitte an eine psychologische Beratungsstelle in Ihrer Nähe. Hilfe bei psychischer Belastung sowie zur Meldung von Hassrede im Internet finden Sie beispielsweise bei der (ZARA Beratungsstelle).

Falls Sie noch Fragen zum Inhalt, Zweck oder Forschungsethik haben oder Interesse an den Ergebnissen haben, wenden Sie sich gerne an Zara Weeke (a12339013@unet.univie.ac.at).

Vielen Dank!

Sie können die Seite nun schließen.

Anhang D

Tabellen

Tabelle D1

Randomisierungsprüfung der soziodemografischen Merkmale nach Bedingungsvergleichen

Merkmal	Test	Statistik	Zusatz (Effekt/Hinweis)
Experimentalbedingungen: HS vs. HS + CS-M vs. HS + CS-eM (*N* = 191)			
Alter	ANOVA	$F(2, 188) = .675$; $p = .510$	Levene: $p = .100$; Welch: $F(2, 124.150) = .605$; $p = .548$
Geschlecht	χ^2 (Pearson)	$\chi^2(4) = 8.213$; $p = .084$	Cramér's V = .147; 3 Zellen (33.3 %) erwartete Häufigkeit < 5 (min = .32)
Wohnort	χ^2 (Pearson)	$\chi^2(2) = .445$; $p = .800$	Cramér's V = .048
Bildungsabschluss	χ^2 (Pearson)	$\chi^2(10) = 19.918$; $p = .030$	Cramér's V = .228; 6 Zellen (33.3 %) erwartete Häufigkeit < 5 (min = .65)
TG-eM	χ^2 (Pearson)	$\chi^2(4) = 6.274$; $p = .180$	Cramér's V = .128; 3 Zellen (33.3 %) erwartete Häufigkeit < 5 (min = 1.30)
CS-Präsenz: HS vs. HS + CS (*N* = 187)			
Alter	ANOVA	$F(1, 185) = .473$; $p = .492$	Levene: $p = .236$; Welch: $F(1, 137.981) = .509$; $p = .477$
Geschlecht	χ^2 (Pearson)	$\chi^2(2) = 6.684$; $p = .035$	Cramér's V = .189; 2 Zellen (33.3 %) erwartete Häufigkeit < 5 (min = .34)
Wohnort	χ^2 (Pearson)	$\chi^2(1) = .056$; $p = .813$	Cramér's V = .017
Bildungsabschluss	χ^2 (Pearson)	$\chi^2(5) = 7.064$; $p = .216$	Cramér's V = .194; 4 Zellen (33.3 %) erwartete Häufigkeit < 5 (min = .67)
TG-eM	χ^2 (Pearson)	$\chi^2(1) = 2.605$; $p = .107$	Cramér's V = .118
CS-Quelle: CS-M vs. CS-eM (*N* = 124)			
Alter	ANOVA	$F(1, 122) = .924$; $p = .338$	Levene: $p = .073$; Welch: $F(1, 121.008) = .927$; $p = .337$
Geschlecht	χ^2 (Pearson)	$\chi^2(1) = .264$; $p = .607$	Cramér's V = .046
Wohnort	χ^2 (Pearson)	$\chi^2(1) = .484$; $p = .487$	Cramér's V = .062
Bildungsabschluss	χ^2 (Pearson)	$\chi^2(5) = 12.841$; $p = .025$	Cramér's V = .322; 5 Zellen (41.7 %) erwartete Häufigkeit < 5 (min = .98)
TG-eM	χ^2 (Pearson)	$\chi^2(1) = .406$; $p = .524$	Cramér's V = .057

Anmerkung. HS = Hate Speech; CS = Counter Speech; CS-M = Counter Speech aus der Mehrheitsgesellschaft; CS-eM = Counter Speech aus der ethnischen Minderheit; TG-eM = ethnische Minderheit. χ^2 = Pearson-Chi-Quadrat-Test; Welch = Welch-ANOVA bei Varianzungleichheit; Levene-Test prüft Varianzhomogenität.

Tabelle D2

Explorative UNIANOVA-Modelle zur wahrgenommenen sozialen Identitätsbedrohung

Modellspezifikation	Effekt	df	*F*	*p*	partielles η²
Einstellung ggü. ethnischen Minderheiten	TG-eM	(1, 176)	22.179	< .001	.112
	CS-Präsenz (HS vs. HS + CS)	(1, 176)	4.057	.046	.023
	CS-Präsenz × TG-eM	(1, 176)	5.409	.021	.030
	Geschlecht	(2, 176)	2.608	.076	.029
Politische Voreinstellung	TG-eM	(1, 176)	26.132	< .001	.129
	CS-Präsenz (HS vs. HS + CS)	(1, 176)	4.682	.032	.026
	CS-Präsenz × TG-eM	(1, 176)	6.032	.015	.033
	Geschlecht	(2, 176)	3.064	.049	.034
Mediennutzung – Beiträge kommentieren	TG-eM	(1, 176)	14.375	< .001	.076
	Geschlecht	(2, 176)	3.212	.043	.035
	Mediennutzung – Beiträge kommentieren	(1, 176)	7.590	.006	.041
	CS-Präsenz × TG-eM	(1, 176)	3.949	.048	.022

Anmerkung. HS = Hate Speech; CS = Counter Speech. TG-M = Mehrheitsgesellschaft; (Referenz: TG-eM = ethnische Minderheit). Berichtet werden Ergebnisse explorativer UNIANOVA-Modelle mit partieller Effektstärke ηp². Levene-Tests prüfen Varianzhomogenität.

Tabelle D3

Explorative GLM-Analysen zur stereotypbezogenen Zuschreibung – Eigenschaftszuschreibung

Modellspezifikation	Effekt	df	*F*	*p*	partielles η²
Einstellung ggü. ethnischen Minderheiten	Alter (zentriert)	(1, 176)	9.195	.003	.050
	Einstellung ggü. ethnischen Minderheiten	(1, 176)	12.131	< .001	.064
Politische Voreinstellung	Alter (zentriert)	(1, 176)	9.569	.002	.052
	Politische Voreinstellung	(1, 176)	13.884	< .001	.073
Mediennutzung – Beiträge kommentieren	Alter (zentriert)	(1, 176)	10.039	.002	.054
	Mediennutzung – Beiträge kommentieren	(1, 176)	5.762	.017	.032
	CS-Präsenz × TG-eM	(1, 176)	4.101	.044	.023

Anmerkung. CS = Counter Speech. TG-eM = ethnische Minderheit (Referenz: TG-M = Mehrheitsgesellschaft). Berichtet werden Ergebnisse explorativer GLM-Modelle mit partieller Effektstärke ηp².

Tabelle D4

Explorative Multivariates GLM-Analysen zur stereotypbezogenen Zuschreibung - Distanz/Vermeidung

Modellspezifikation	Effekt	df	*F*	*p*	partielles η²
Einstellung ggü. ethnischen Minderheiten	Einstellung ggü. ethnischen Minderheiten	(1, 176)	18.472	< .001	.095
	CS-Präsenz × Einstellung	(1, 176)	4.412	.037	.024
	CS-Präsenz × TG-eM × Einstellung	(1, 176)	4.288	.040	.024
Politische Voreinstellung	Politische Voreinstellung	(1, 176)	4.441	.037	.025
Mediennutzung – Beiträge kommentieren	CS-Präsenz × Mediennutzung – Beiträge kommentieren	(1, 176)	6.513	.012	.036
	TG-eM × Mediennutzung – Beiträge kommentieren	(1, 176)	8.063	.005	.044
	CS-Präsenz × TG-eM × Mediennutzung – Beiträge kommentieren	(1, 176)	8.770	.003	.047

Anmerkung. CS = Counter Speech. TG-eM = ethnische Minderheit (Referenz: TG-M = Mehrheitsgesellschaft). Berichtet werden Ergebnisse explorativer GLM-Modelle mit partieller Effektstärke ηp².

Tabelle D5

Explorative UNIANOVA-Analysen zur Interventionsbereitschaft – Schweregradwahrnehmung

Effekt	df	*F*	*p*	partielles η²
TG-eM	(1, 180)	4.829	.029	.026
Alter	(1, 180)	13.988	< .001	.072
CS-Präsenz	(1, 180)	.215	.644	–
CS-Präsenz × TG-eM	(1, 180)	3.072	.081	.017

Anmerkung. CS = Counter Speech. TG-eM = ethnische Minderheit (Referenz: TG-M = Mehrheitsgesellschaft). Berichtet werden Ergebnisse explorativer UNIANOVA-Modelle mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

Tabelle D6

Explorative UNIANOVA-Analysen zur Interventionsbereitschaft – Verantwortungsgefühl

Effekt	df	*F*	*p*
Gesamtmodell	(6, 180)	.742	.616
CS-Präsenz	(1, 180)	.000	.990
TG-eM	(1, 180)	.004	.953
Alter (zentriert)	(1, 180)	.863	.354

Anmerkung. CS = Counter Speech. TG-eM = ethnische Minderheit (Referenz: TG-M = Mehrheitsgesellschaft). Berichtet werden Ergebnisse explorativer UNIANOVA-Modelle mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.

Tabelle D7

Explorative UNIANOVA-Analysen zur Interventionsbereitschaft – Interventionsabsicht (Einzelitem)

Effekt	**df**	***F***	***p***	**partielles η²**
Gesamtmodell	(6, 180)	1.211	.302	–
CS-Präsenz	(1, 180)	2.001	.159	–
TG-eM	(1, 180)	1.956	.164	.011
Geschlecht	(2,180)	.256	.774	.003
Alter (zentriert)	(1, 180)	2.806	.096	.015

Anmerkung. CS = Counter Speech. TG-eM = ethnische Minderheit (Referenz: TG-M = Mehrheitsgesellschaft). Berichtet werden Ergebnisse explorativer UNIANOVA-Modelle mit partieller Effektstärke ηp². Levene-Test prüft Varianzhomogenität.