MUNI FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

Schmid, U. K., Greipl, S., & Rieger, D. (2026). Subtle cues, major shifts: Eye-tracking insights on how fear and humor legitimize hostility on social media. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 20*(3), Article 7. https://doi.org/10.5817/CP2026-3-7

Subtle Cues, Major Shifts: Eye-Tracking Insights on How Fear and Humor Legitimize Hostility on Social Media

Ursula Kristin Schmid, Simon Greipl, & Diana Rieger

Department of Media and Communication, LMU Munich, Germany

Abstract

Expressions of hostility are widespread online, occurring not only in explicit but even more so in implicit forms. In this study, we trace the subtle yet lasting harmful potential of two prevalent implicit forms of online hostility—fear speech and derogatory humor—by examining their perceived acceptability and legitimacy. We argue that both forms are more compatible with mainstream discourse than explicit hate and function as manifestations of symbolic violence, providing justifications that legitimize hostility and normalize extreme positions. In an experimental study with eye-tracking measures (N = 141) among predominantly young social media users in Germany (age range: 18-65; M = 23.33, SD = 8.02), we examined users' attention to and perception of fear speech and derogatory humor on social media compared to explicit hostility. The results reveal that fear speech and derogatory humor attract similar or even greater attention than explicit hostility but differ in perception: both increase acceptance, yet fear speech elicits a heightened perceived potential to incite violence. Our findings provide insights into subtle legitimization processes and highlight broader implications for the normalization of extreme viewpoints.

Keywords: hostility; fear speech; derogatory humor; social media; eye-tracking

Editorial Record

First submission received:
March 13, 2025

Revisions received:
October 15, 2025
April 2, 2026

Accepted for publication:
April 2, 2026

Editor in charge:
Lenka Dedkova

Introduction

Online environments are widely recognized for fostering toxic and hateful discourses, with extensive literature documenting explicit forms of hostility, such as hate speech (Hawdon et al., 2017). However, recent research highlights that implicit forms of hostility—such as the portrayal of threatening "others" (referred to as fear speech; Greipl et al., 2024) and derogatory humor as a form of humorous hate speech (Schmid, 2025)—may be even more prevalent. Importantly, implicit hostility operates differently from explicit forms; its covert nature often allows it to be justified and accepted by recipients, while potentially amplifying harm (Recuero, 2024). This often manifests in humorous reframing through memes and pop-cultural references, as well as fear-driven narratives, for instance about migrant criminality (Fielitz, et al., 2024). Particularly around the latter topic, Germany and other Western democracies have seen a rise in exclusionary discourse in recent years. Right-wing actors such as the AfD (German right-wing party "Alternative für Deutschland") have employed irony, fear-based rhetoric, and cultural references to express hostility in ways that avoid explicit hate yet resonate broadly. These discursive strategies reflect core

dynamics of communicating prejudice through socially acceptable justifications, as exemplified in the justification–suppression model (JSM; Crandall & Eshleman, 2003).

Fear speech and derogatory humor, as particularly subtle yet pervasive forms of power enactment, illustrate what Recuero (2015, 2024) identifies as symbolic violence in social media environments. Operating through language, imagery, and implicit communication, symbolic violence reinforces social hierarchies and normalizes systems of exclusion and domination (Bourdieu, 1991; Žižek, 2007). Symbolic violence is integral to legitimizing discriminatory narratives, as it frames devaluation and even dehumanization as, for instance, humorous or threatening—and thus potentially socially acceptable. This process can reshape societal norms, making exclusionary ideologies appear more palatable (Billig, 2001; Rothut et al., 2024). Particularly by offering plausible justifications—such as framing a message as 'just a joke' or as a rational response to a perceived threat—such implicit forms of hostility can make prejudiced expressions appear more socially acceptable and diminish the perceived severity of others' hostile messages (Hodson & MacInnis, 2016). While explicit hostility is typically tied to overt dominance motives and appeals primarily to individuals who strongly endorse hierarchical social structures (Bilewicz et al., 2017), implicit expressions can resonate with a broader audience that may also hold internalized beliefs in social inequality, but feel the need to suppress the expression or even the endorsement of these prejudices (Crandall & Eshleman, 2003). Consequently, symbolic violence or implicit hostility reduces the stigma associated with hostile expressions of ingroup-outgroup dynamics, embedding these ideas deeper into the collective consciousness. Over time, this not only reduces resistance to overt violence but also integrates exclusionary narratives into mainstream discourse, solidifying them as part of the cultural status quo (Papacharissi, 2015).

As social media increasingly shape public discourse, identifying and addressing symbolic violence in subtle hostilities is challenging, yet essential for preventing legitimization and normalization of harmful ideologies. In this experimental study, we aim to trace the foundational processes, including (1) audience's visual attention to the message and (2) their subjective perception, that may facilitate legitimization and subtle normative shifts at the individual level. For this purpose, we combine eye-tracking and survey data in a laboratory experiment to compare the reception of hostility in explicit forms with more implicit forms, namely fear speech and derogatory humor. Our findings illustrate how explicit, implicit, and non-hateful content not only captures recipients' attention in distinct ways but is also perceived differently across critical dimensions of perceived legitimacy and justification, such as social acceptability, hostility, and perceived potential to incite violence.

(Strategic) Legitimization on Social Media

In recent years, democratic countries have witnessed a shift as extreme political views, including far-right ideologies, increasingly permeate the broader public (Mudde, 2019). From a communication science perspective, Rothut et al. (2024) describe this process as mainstreaming of extreme views; an iterative process driven by both unintentional factors, such as broader societal shifts, and intentional strategies employed by extreme actors. On the surface, these strategies involve content positioning. However, it also entails reaching a broad audience by, strategically speaking, increasing the audience's susceptibility to the content or by making it appear significant. Arguably, social media platforms are effective in that domain too, as they reward emotional content with high visibility (Schulze et al., 2024), which—when accompanied by perceived acceptability—forms a critical pathway toward its legitimization.

Legitimization is a central step in the mainstreaming process. Particularly discussed from critical discourse analysis perspectives, it refers to the extent to which discourses align with the social norms and values of a society at a given time (Reyes, 2011; van Leeuwen, 2007; van Leeuwen & Wodak, 1999). Complementing this, *delegitimization*, particularly relevant in the context of intergroup conflicts, has been examined extensively. Bar-Tal and Hammack (2012) describe delegitimization as a normative process that stems from the division of societies into in- and outgroups, wherein members of the outgroup are rhetorically assigned a diminished moral and existential worth. Importantly, the two concepts of de-/legitimization are both closely related to power and dominance, and viewed as a key cognitive mechanism in humans' tendency to maintain hierarchical social structures (Sidanius & Pratto, 1999). Through (de-)legitimization, often also referred to as justification, (political) actors not only seek to secure immediate approval and support from their audience but also to obtain or sustain power (Bar-Tal & Hammack, 2012; Reyes, 2011). To achieve this goal, *legitimization strategies* are (linguistic) tools that these actors use to justify and legitimize their ideological positions (Reyes, 2011), often by devaluing or ridiculing threatful others. Recently found in internet memes (Davis et al., 2016; Ross & Rivers, 2017) or politicians' social media posts (Koivukoski et al., 2025; Recuero & Soares, 2022), legitimization strategies primarily aim to increase the tacit

agreement and acceptance of the audience (Recuero, 2024). The potential behind this strategy is also highlighted by the JSM (Crandall & Eshleman, 2003), which posits that many individuals harbor prejudiced attitudes, but social norms often suppress their overt expression. Consequently, individuals are more likely to articulate such views when they encounter justifications that render biased statements socially acceptable.

Legitimization of Hostility via Fear Speech and Derogatory Humor

Legitimization, as it is based on power dynamics, is best understood within hostile online communication where social (out-)groups are targeted. A common strategy to mask and thereby justify norm-violating content is the use of implicit rather than explicit forms of hostility (Ben-David & Matamoros-Fernandez, 2016) so that hostility can be perceived as benign (Schmid & Greipl, 2025), while still effectively drawing attention. Two prevalent forms of such implicit hostility used by both political and non-institutional actors as well as ordinary users are *fear speech* and *derogatory humor*.

While fear speech "portrays a particular entity, e.g., a group or an institution, as inherently and/or imminently harmful on a cultural, societal, or existential level" (Greipl et al., 2024, p. 10), derogatory humor—also referred to as humorous hate speech (Schmid, 2025; Schmid & Greipl, 2025; Yeon & Lee, 2021)—involves the denigration, belittlement, and defamation of others in even more subtle ways, under the guise of humor (Ford & Ferguson, 2004).

By drawing on both fear and humor, communicators can justify hostility, expressing and normalizing prejudice in ways that reduce or even avoid overt social rejection. While communicating fear has long been discussed as the main potent strategy in this regard (Reyes, 2011; van Leeuwen & Wodak, 1999; Wodak, 2021), the legitimization potential of humorously framed content receives increasing scholarly attention in various contexts, including its role in promoting online hostility by framing prejudiced expressions as harmless fun (Crilley & Chatterje-Doody, 2021; Hodson & MacInnis, 2016; Recuero & Soares, 2022; Ross & Rivers, 2017). Although fear speech and derogatory humor represent distinct communication strategies, they share key communicative as well as psychological commonalities. On the communicative level, both convey hostility in a veiled and implicit manner by drawing on ingroup-outgroup narratives reinforcing underlying group-dominance motives and power hierarchies. On the psychological level, both rely on emotions that align more closely with mainstream sensibilities than explicit hate, while engaging advanced cognitive and affective processing pathways.

A central discursive practice in the context of strategic legitimization is (rhetorically) constructing a 'them-group' that is distinct from the 'us-group', including the speaker and the audience (Oddo, 2011; Reyes, 2011; van Leeuwen & Wodak, 1999). Both derogatory humor and fear speech communicate outgroup prejudices while simultaneously strengthening ingroup cohesion (Ford & Ferguson, 2004; Hodson et al., 2010; Meiering et al., 2018). By maintaining the boundary between ingroup and outgroup, both perpetuate and legitimize power relations.

Within fear speech, outgroups (the 'them-group'), such as migrants arriving in Germany, are typically portrayed as threatening entities (Greipl et al., 2024; Saha et al., 2021). This portrayal reinforces images of the outgroup as the (dangerous) enemy (Marcks & Pawelz, 2022), while promoting identification with the supposedly disadvantaged ingroup (Freiheit & Zick, 2022; Meiering et al., 2018). Fear speech amplifies perceived threats and provides a moral rationale for oppression, framing domination as necessary for order and security. This rationalization might even lead to the legitimization of violence in the sense of self-defense (Buyse, 2014; Sayimer & Derman, 2017). Similarly, Bar-Tal and Hammack (2012) outline how rhetorical delegitimization in intergroup conflicts can lead to the justification of ingroup immoral acts, including the use of violence.

Humor operates similarly by conveying a dualistic 'us-versus-them' thinking: aggressive and derogatory forms of humor promote stereotypes about the outgroup and rhetorically support a logic of exclusion and 'othering' (Weaver, 2011). In doing so, derogatory humor marks outgroups as legitimate targets of ridicule and exclusion (Hodson & MacInnis, 2016). It is deeply entwined with power relations (Davies & Ilott, 2018), frequently reinforcing ingroup superiority (Lintott, 2016), and helping to maintain racialized hierarchies (Sakki & Castrén, 2022; Weaver, 2011).

On a psychological level, using emotional language is considered one of the most important legitimization strategies, particularly in political communication (Oddo, 2011). Although fear speech and derogatory humor operate at different ends at the emotional valence spectrum—fear speech on the negative, and humor on the positive end—both rely on affective appeals that resonate with mainstream audiences more readily than explicit hate. Fear evokes anxiety and vulnerability (Greipl et al., 2024; Klein, 2021), echoing prevalent emotions and

addressing concerns especially during times of crises. As such, communicating fear has been found frequently to be an effective strategy for justifying hostility and (political) violence (Albertson & Gadarian, 2015; Oddo, 2011). Humor conveys hostility indirectly by mocking other groups and presenting ridicule as entertainment. In doing so, derogatory humor primes the expectation and perception of levity (Ford et al., 2008). This levity is particularly highlighted on social media platforms where derogatory humor is frequently combined with (pop-) cultural elements, for instance within memes (Schmitt et al., 2020). Disguised as entertaining content, derogatory humor offers communicators the benefit of trivializing the hostile message (Chovanec, 2021), as it could be interpreted as 'just fun' (Billig, 2001), allowing communicators to distance themselves from the message if challenged (Matamoros-Fernández et al., 2023; Pérez, 2013).

The emotions elicited by fear (speech) and (derogatory) humor also engage deeper information processing that heightens attention and cognitive/affective involvement. Empirical evidence suggests that messages featuring fear and humor stimulate increased attention and information seeking. Fear sensitizes individuals to threatening cues and amplifies risk perception (Bar-Tal, 2013), while anxiety—a common consequence of sustained fear—directs attention toward the environment and motivates information acquisition (see Affective Intelligence Theory, AIT; Marcus et al., 2000). Humor elicits positive affect, enhances memory and recall (Borah, 2016; Coronel et al., 2021), and demands additional cognitive processing to resolve incongruities (Suls, 1972), particularly when humor and hostility intertwine (Schmid, 2025).

Taken together, these key features illustrate how fear speech and derogatory humor represent highly functional legitimization strategies in the sense of the JSM (Crandall & Eshleman, 2003). On a continuum of message extremity, fear speech and derogatory humor may occupy a strategic middle ground—provocative enough to attract attention but not so extreme as to trigger disengagement. By drawing on 'justified' emotions and in-group/out-group narratives, they construct socially acceptable frames that may reduce both internal and external inhibitions against prejudice. This dynamic not only facilitates the expression of dominance and power over marginalized groups, but also increases the likelihood that audiences tacitly accept hostile content within mainstream social media discourse.

The Reception Process in the Legitimization of Hostility

Bringing hostile or extremist positions into the broader public involves to increase the visibility of such content within mainstream environments, and to enhance and exploit recipients' susceptibilities (Rothut et al., 2024, 2026). We argue that to fulfill these prerequisites, and in consequence for legitimization strategies to work, the initial moment of reception is critical. Although extreme views are normalized gradually, these initial interpretations may facilitate the legitimization of (more extreme forms of) hostility and contribute to the normalization of violence as well as the shifting of social norms towards the extremes (Rothut et al., 2024; Wodak, 2021). Social media users' first encounters with such speech should therefore reflect an important initial step in any mainstreaming process.

The initial reception of such content relies on two fundamental components that jointly determine the effectiveness of a legitimation strategy. First, the extent and quality of attention the content attracts—reflecting the communicator's success in strategically positioning the message. The second involves the audience's susceptibility, manifested in how recipients perceive and evaluate the content (Rothut et al., 2024). We do not assume a simple linear relationship between visibility and perception; rather, both components must coincide to become effective mechanisms of legitimization. High levels of attention alone are insufficient if the content simultaneously provokes shock, is perceived as unjustified or inappropriate, and thereby elicits negative perceptions (Rieger et al., 2013). Conversely, a benign or harmless perception of content is ineffective if the content fails to attract sufficient attention. The critical point lies at the intersection: elevated attention that remains unchallenged—that is, attention not accompanied by moral rejection or negative perceptions. It is this (balancing) dynamic of visibility without resistance that enables hostility to enter mainstream discourse and gradually gain legitimacy. In the following sections, we examine the extent to which implicit hostility, manifested through derogatory humor and fear speech, fulfills these two conditions in comparison to explicit hostility.

Visual Attention to Implicit and Explicit Hostility

Considering visual attention as a first necessary step in the legitimization process, research has shown that hostile comments tend to attract users' attention, more so than civil ones (Santana & Hopp, 2022). Hostility itself may act as a salient cue, capturing attention through its social and moral relevance. Explicit hostility, characterized by overt

aggression and prejudices, should therefore be more likely to draw immediate attention. Implicit forms, such as fear speech or derogatory humor, tend to obscure their hostile intent and may initially go unnoticed (e.g., for humorous hate speech; Schmid, 2025).

At the same time, the relationship between hostility and attention is not linear. Qualitative and quantitative observations showed that many users tend to be disengaged or deterred by posts with overtly aggressive tones (Schmid et al., 2024) or extremist messages (Rieger et al., 2013). This suggests that explicit hostility may capture attention momentarily but fail to sustain it, because attention is not only a function of intensity, but also incongruity and expectation violation (Knobloch-Westerwick, 2014; Lang, 2000). Veiled hostility that embeds antagonistic meaning in familiar affective or cultural frames may have at least a long-term advantage by striking this balance. It should appear relevant while avoiding the perceptual overload or moral rejection often triggered by explicit hate. In addition, social media affordances reward content that sustains engagement without breaching platform norms or social desirability thresholds (Kakavand, 2024). As such, implicit hostility may not be attentionally disadvantaged but rather optimized for visibility in algorithmically curated, socially moderated spaces. Based on this background, it is unclear whether explicit or implicit forms of hostility gain more attention when embedded in a social media post. We therefore ask:

RQ1: Are there differences in social media users' visual attention to explicit and implicit (i.e., fear speech and derogatory humor) hostility?

Perception of Implicit and Explicit Hostility

While visibility marks the entry point of hostile content into users' cognitions, their subsequent interpretation of that content, reflected in the actual perception, constitutes the second critical step in the legitimization process (Rothut et al., 2024). Content that attracts attention does not necessarily influence attitudes; what matters equally is how it is perceived and evaluated. Perception determines whether hostility is recognized and rejected or reinterpreted as acceptable when embedded in justifying frames such as humor or fear.

Empirical research suggests that implicit forms of hostility are generally perceived as less extreme and less morally objectionable than their explicit counterparts (Schmid et al., 2024). This can be attributed to the ambiguous and indirect presentation of hostility, which allows audiences to downplay or rationalize its intent (Feischmidt & Hervik, 2015). When antagonism is expressed through fear-based or humorous cues, recipients are less likely to identify its extreme underpinnings and may instead interpret it as justified concern or harmless entertainment. Moreover, the affective resonance of fear speech and derogatory humor allows them to align more closely with contemporary cultural norms, making these forms of hostility appear more tolerable. For instance, social media users perceive humorous forms of hate speech more socially acceptable, but less hostile than non-humorous forms (Schmid & Greipl, 2025). Taken together, we assume that implicit forms of hostility are more likely to elicit perceptions of justification and acceptability, while being less likely to elicit perceptions of hostility and harmfulness.

H1: Social media users perceive implicit hostility (i.e., fear speech and derogatory humor) as more justified and less harmful than explicit hostility.

Cavalier Humor Beliefs as Legitimizing Myth

Personality and attitudinal characteristics of the audience may further determine whether the content falls within their individual range of (perceived) legitimacy or not. Specifically, dominance-orientation motives, as represented in both fear speech and derogatory humor, do not appeal equally to individuals. One personality trait linked to dominance orientation is *cavalier humor beliefs* (CHBs), capturing "a lighthearted, less serious, uncritical, and nonchalant mindset toward humor generally" (Hodson et al., 2010, p. 663). Research indicates that individuals with higher levels of cavalier humor beliefs are more likely to perceive derogatory humor as harmless and 'just fun' (Buie et al., 2022; Hodson et al., 2010; Prusaczyk & Hodson, 2020; Schmid, 2025).

Beyond representing a lighthearted attitude toward (derogatory) humor, CHBs have been linked with a generalized negativity toward outgroups (Hodson et al., 2010; Hodson & MacInnis, 2016). Within the framework of Social Dominance Theory (Sidanius & Pratto, 1999), they have been conceptualized as *legitimizing myths* (Hodson et al., 2010; Hodson & MacInnis, 2016)—beliefs that facilitate the expression of dominance motives and justify the maintenance of social inequality. Whereas traditional legitimizing myths, such as overt ideologies of racial

superiority, have declined in many Western societies, CHBs represent a more subtle and socially acceptable mechanism that masks prejudice under the guise of humor. Viewed through the JSM (Crandall & Eshleman, 2003), CHBs function as ideological justifications that lower inhibition thresholds and recast prejudiced discourse as harmless amusement. In doing so, they reduce both internal restraints and external sanctions, contributing to the legitimization of subtle hostility.

While CHBs were initially conceptualized as a humor-specific disposition, we suggest that they may also capture a broader sensitivity to group-based status dynamics. From this perspective, CHBs not only legitimize status-raising at the expense of others through humor but may also reduce resistance to messages emphasizing status threats, as conveyed in fear speech. Both forms thus engage a shared propensity to rationalize hostility when it is linked to intergroup hierarchy. This theoretical extension situates CHBs within the broader logic of legitimizing myths and connects them to the JSM, offering a rationale for why CHBs might moderate perceptions of both derogatory humor and fear speech.

RQ2: Do cavalier humor beliefs moderate the perceptions of implicit hostility (i.e., fear speech and derogatory humor)?

Methods

Design and Participants

To answer our research questions, we conducted a between-subject laboratory experiment in which participants were exposed to social media posts containing either implicit hostility (fear speech or derogatory humor), explicit hostility, or a neutral control message. Participants' visual attention—representing the visibility component of the legitimization process—was assessed via eye-tracking (RQ1). Their perceptions of the post content (H1) and individual dispositions, particularly CHBs, were measured through post-exposure questionnaires (RQ2). The study was conducted in December 2023 in the premises of the Department of Media and Communication of LMU Munich and mainly included students, who are regularly exposed to similar social media content. Participants were recruited from among students attending a university seminar on the study's broader topic. Recruitment took place either on university premises, during seminar sessions, or via personal networks such as friends and acquaintances. During recruitment, participants were informed about the general subject of the study and the methodological approach. However, no specific details about the stimuli were disclosed to avoid biasing their responses. The study was reviewed and approved by the Ethics Committee of the Faculty of Social Sciences at LMU Munich (approval number: 25-07).

After the deletion of the data of 15 participants according to our quality criteria (technical problems in the lab session; no recall of the relevant social media post), 141 people remained in the final sample (age range: 18-65; M = 23.33, SD = 8.02, 53% women, 43% with migration background, 96% higher education [baccalaureate or higher], 86% students). At the highest usage level (several times a day), messenger services (e.g., WhatsApp, Telegram, Signal) were used by 92.2% of participants, social media platforms (e.g., TikTok, Instagram, Facebook) by 70.9%, video and livestreaming platforms (e.g., YouTube, Twitch) by 37.6%, video games by 5%, gaming platforms and forums (e.g., Discord, Steam, Teamspeak) by 8.5%, and other online platforms by 4.3%. The study was sufficiently powered to detect small-to-medium main effects ($f^2 \approx .06$, $1-\beta = .80$, $\alpha = .05$), which was the main objective for RQ1 and H1.

Stimulus Material and Experimental Groups

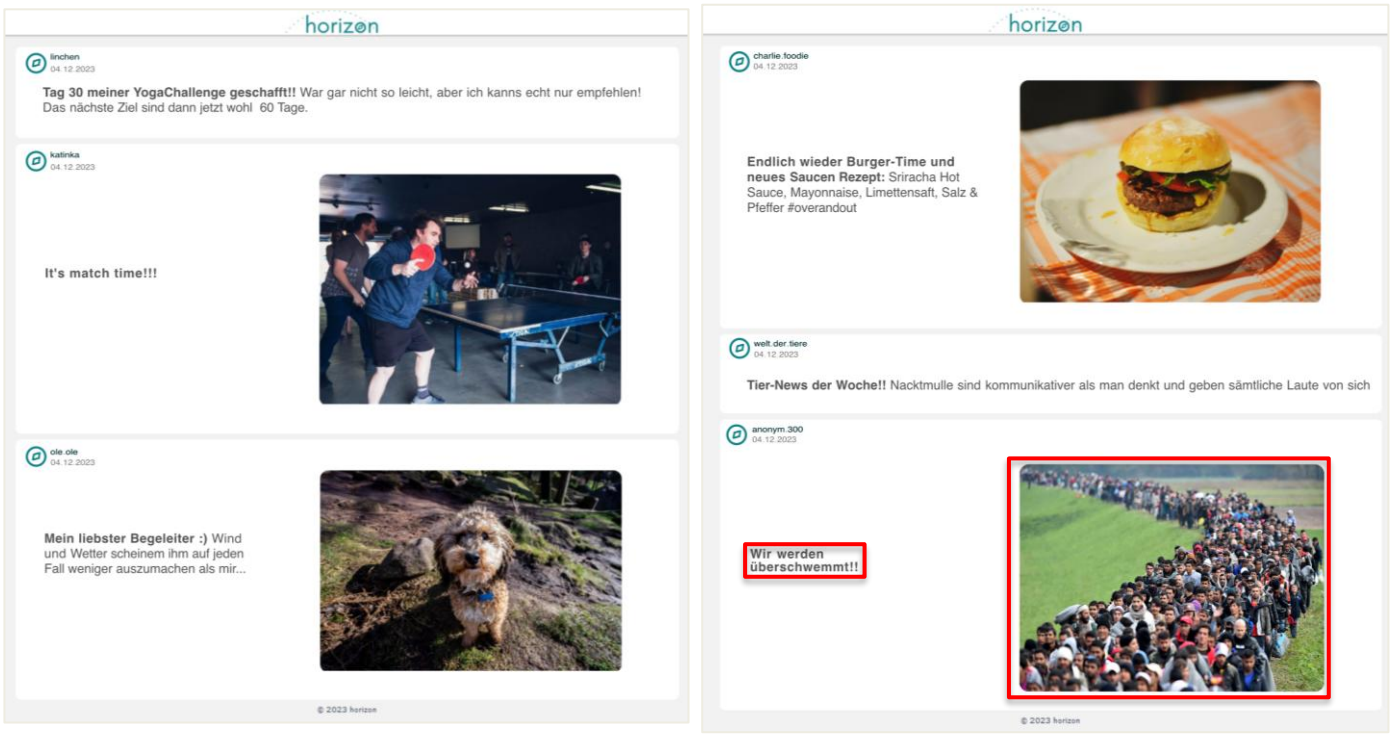
The stimulus material consisted of a social media posting communicating anti-migrant attitudes, which are prevalent in current public discourse on social media (Nann et al., 2024), often communicated alongside fear or humor to legitimize and mainstream radical political viewpoints (McSwiney & Sengul, 2024; Sayimer & Derman, 2017), particularly in visuals (Rothut et al., 2024, 2026; Schmid et al., 2025).

Participants were shown two screenshots of the feed of a fictitious social media platform, with the second screenshot containing one post that expressed hostility toward migrants arriving in Germany (see Figure 1). The independent variable (IV) was realized in four levels: Within this post, hostility was expressed either (1) explicitly or implicitly by means of (2) fear speech or (3) derogatory humor. For comparison, (4) a neutral control condition was integrated. Importantly, the levels of the IV were realized by altering the text of the post while keeping the

visual motive constant (Figure 2). To ensure optimal balance between external and internal validity, word length was kept (approximately) the same, but wordings were adjusted across the four versions. Accordingly, all IVs conveyed an anti-migrant sentiment with varying emphases, such as on explicit hostility ("All scum!!"), fear-based elements ("We are flooded!!"), or humor elements ("Prison Hiking Day!!"). This approach aligns with previous studies that compared humorous and explicit hateful stimuli (Schmid, 2025; Schmid & Greipl, 2025; Yeon & Lee, 2021). To ensure that the effects were not driven by the specifics of a single stimulus motive, two versions of the target post with similar pictures were employed. Both focused on migrants and contained the same core message (e.g. fear speech in variant one: "We are flooded!" and fear speech in variant two: "We are overrun!") but differed in their visual motive (see Figure 2).

Figure 1. *Stimulus Material.*

Screenshot 1 **Screenshot 2**



Note. Both screenshots were shown for 20 seconds. The hostile post can be found at the bottom of screenshot 2, which shows implicit hostility through fear speech in post motive 1. Framed in red are the key areas of interest (text and picture of the relevant post).

Participants were randomly assigned to one of the eight conditions (four hostility types with each two stimulus motives). Experimenters were blind to the condition. As no systematic differences in outcome variables were observed between the two stimulus motives (all Fs < 1.35, ps > .24)[1], responses were aggregated by hostility type, resulting in four groups for comparison. After aggregation, there were no significant differences across the experimental groups ($N_{\text{explicit hostility}}$ = 43; $N_{\text{fear speech}}$ = 36; $N_{\text{derogatory humor}}$ = 35; N_{control} = 27) regarding age ($F(3, 137) = 0.41$, $p = .748$) and migration background ($\chi^2(6) = 10.133$, $p = .119$). The distribution of sex was slightly skewed; however, when $N = 2$ diverse participants were excluded (only for the structural comparison), there were no differences between the experimental groups ($\chi^2(3) = 5.98$, $p = .11$). Finally, a MANOVA including all six media usage categories (messenger services, social media platforms, video- and livestreaming platforms, video games, gaming platforms and forums, and other platforms) indicated no significant differences in overall social and other online media consumption patterns across experimental groups (*Pillai's Trace* = 0.15, $F(18, 405) = 1.19$, $p = .267$), indicating that randomization was successful for these variables.

Figure 2. *Experimental Variation of the Stimulus Material.*

Explicit Hostility	Fear Speech	Derogatory Humor	Neutral	Stimulus Motive
				Motive 1
All scum!! *(Original German Version: Alles Abschaum!!)*	We are flooded!! *(Original German Version: Wir werden überschwemmt!!)*	Prison Hiking Day!! *(Original German Version: JVA Wandertag!!)*	Situation at the border!! *(Original German Version: Lage an der Grenze!!)*	
				Motive 2
All waste!! *(Original German Version: Alles Abfall!!)*	We are overrun!! *(Original German Version: Wir werden überrannt!!)*	Wacken 2045!! *(Original German Version: Wacken 2045!!)*	Picture of the day!! *(Original German Version: Bild des Tages!!)*	

Note. Participants were randomly assigned to one of the 8 conditions. For analytical purposes, the dependent variables for Stimulus Motive 1 and 2 were aggregated within each experimental condition, as no systematic differences were observed.

Procedure and Measures

Upon providing verbal consent to participate in the study, participants were guided into the laboratory where they gave written consent and were given general information about the study and the eye-tracking procedure. Participants were informed that they could decline participation at any time for any or no reason, especially if they felt uncomfortable with questions or materials.

Each session was conducted in the same room and with unchanged surroundings (e.g., same lighting conditions). To ensure optimal eye-tracking results, participants were seated approximately 60 cm from the eye-tracker (Tobii Pro Fusion), a panel which was attached to the bottom edge of a 24" screen with a resolution of 1920 x 1200 pixels. Eye movements were tracked from both eyes (binocular tracking) at a sampling frequency of 120 Hz. Prior to the experiment, each participant completed a nine-point calibration task, which was repeated as necessary to ensure accurate adjustment of the eye-tracker. During calibration, participants were instructed to follow a dot displayed on the desktop screen, thus ensuring precise eye movement tracking. While no strict numerical exclusion criteria were applied, calibration quality was assessed visually during this setup phase based on the calibration output provided by the eye-tracking software, and recalibration was performed whenever accuracy was deemed insufficient by the experimenter. The final sample (N = 141) achieved high spatial accuracy ($M_{accuracy}$ = 0.64°, SD = 0.35°) and minimal data loss (M = 4.29%). Sensitivity analyses confirmed that the reported effects remained robust when employing strict exclusion thresholds (e.g., accuracy < 1.0°).

Participants then completed the first part of a questionnaire, which collected sociodemographic information (including age, sex and migration background). Additionally, *cavalier humor beliefs* were assessed using a 5-point scale (*do not agree at all* to *fully agree*), based on Hodson et al.'s (2010) six-item scale, translated into German (Schmid, 2025). An example item reads, "jokes are simply fun" (α = .78, M = 2.62, SD = 0.76).

Following the first part of the questionnaire, participants read an introductory text informing them that they would be viewing a part of a social media platform on the next pages. They were then shown the two screenshots containing the stimulus material, each displayed for 20 seconds. During this period, participants' eye movements

were tracked, focusing on their attention to specific areas of the stimulus (areas of interest, AOIs). We recorded the *total dwell time* and the *number of fixations* on the AOIs. Total dwell time refers to the aggregate time or duration a gaze lingers on a specific area of interest, while the number of fixations counts the brief pauses of the eye when it focuses on a particular point, with both metrics representing key indicators of visual attention (King et al., 2019). Using Tobii Pro Lab's I-VT Gaze Filter, any eye movement with an angular velocity below 30°/s is considered a fixation, with a minimum fixation duration of 60 milliseconds. As AOIs, we defined the (1) text of the hostile social media post that varied between the experimental conditions as well as (2) the picture of this post (see Figure 1 and Figure 2).

After viewing the screenshots, participants were directed to the second part of the online questionnaire and were first asked to recall the content of the social media posts, with *correct recall of the targeted post* serving as a quality criterion to be included within our analyses. Next, participants' *subjective perceptions of the hostile post* were measured on 5-point scales (*do not agree at all* to *fully agree*). We exploratively examined multiple different dimensions reflecting both the justification and suppression components outlined in the JSM (Crandall & Eshleman, 2003). Participants rated how humorous (*humorous*, *amusing*, r = .87; M = 1.20; SD = 0.55), *socially acceptable* (M = 2.11; SD = 1.22), *interesting* (M = 2.18; SD = 1.21), and *hostile* (*hurtful*, *offensive*, *discriminatory*, *hateful*, *aggressive*, α = .94, M = 3.61; SD = 1.21) they perceived the post to be. Finally, they evaluated the post's *potential to incite violence against migrants* (M = 3.91; SD = 1.12). Following the JSM (Crandall & Eshleman, 2003), perceptions of humor, interest, and social acceptability can be interpreted as justificatory mechanisms that downplay harm and increase tolerance toward hostility (e.g., for humor; Hodson & MacInnis, 2016), whereas perceptions of hostility and violent potential represent inhibitory appraisals that may trigger moral or social restraint. Together, these measures capture the perceptual processes through which implicit and explicit hostility may differentially become legitimized in online discourse.

Results

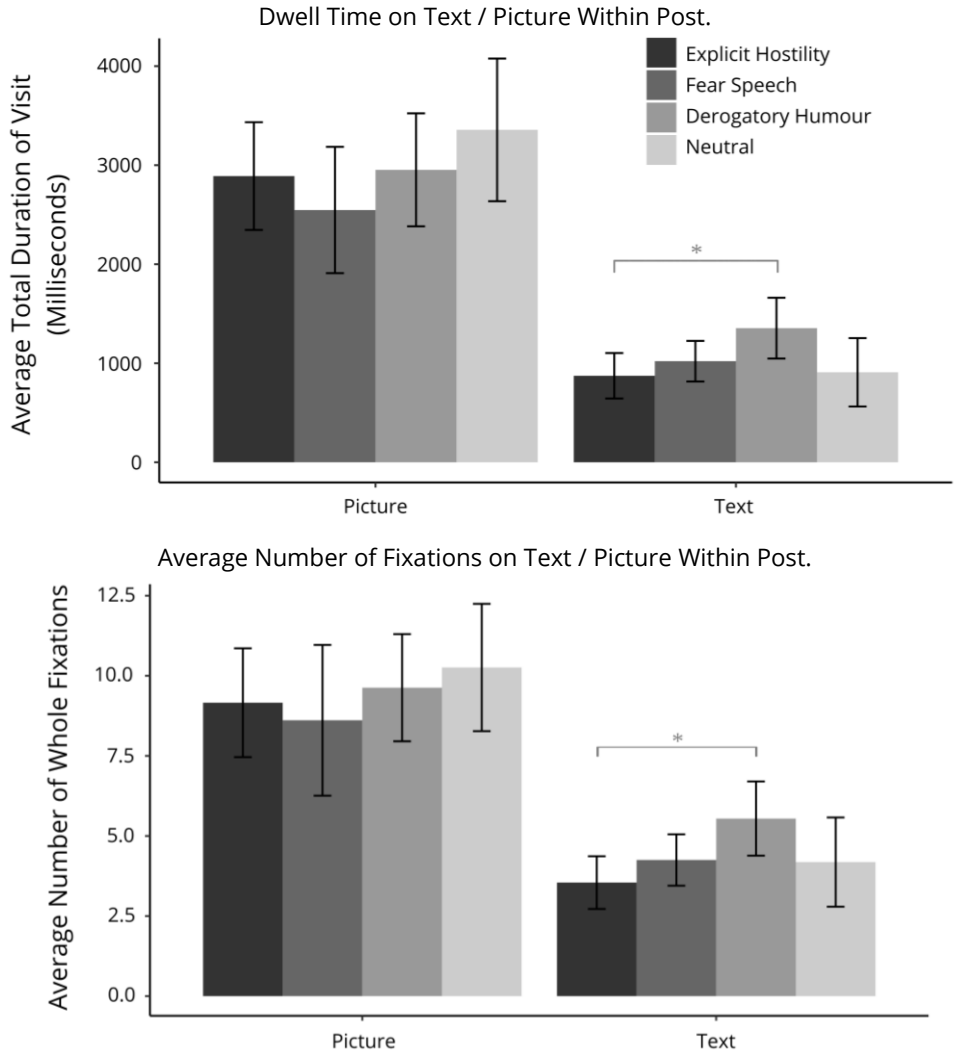
Attention to the Posts

Our first research interest was to compare the visual attention that social media users pay to posts featuring different forms of hostility (RQ1). For this purpose, we calculated separate analyses of variance (ANOVAs) for total dwell time and the number of fixations on the AOIs (text to post and picture to post) to compare them between the experimental conditions (explicit hostility, fear speech, derogatory humor, neutral control condition).

For the text elements, we found significant group differences between participants' dwell time (average total duration of visit; $F(3, 137) = 2.90$, $p = .037$, $\eta^2 = .06$), and the number of fixations ($F(3, 137) = 2.97$, $p = .034$, $\eta^2 = .06$). Participants in the condition with derogatory humor spent significantly more time (in milliseconds) on the text element (derogatory humor: M = 1354.29, SD = 893.99; explicit hostility: M = 859.42, SD = 757.59; p = .031; family-wise error corrected, Holm) and fixated it more often (derogatory humor: M = 5.54, SD = 3.37; explicit hostility: M = 3.53, SD = 2.74; p = .019; family-wise error corrected, Holm) than participants who saw the explicit hostility. There were no significant differences between fear speech (dwell time on text element: M = 1020.94, SD = 605.86; fixations on text element: M = 4.25, SD = 2.37) and the other experimental conditions (e.g. in the neutral condition, dwell time on text element: M = 908.59, SD = 872.71; fixations on text element: M = 4.19; SD = 3.52), nor any differences in participants' attention to the picture of the social media posts (see Figure 3). Answering RQ1 on differences in visual attention to explicit versus implicit forms of hostility, we find a significant differentiation only with respect to derogatory humor.

We examined in preliminary tests whether the two steps in the legitimation process—attention to the post and the perception dimensions—are independent of one another. Theoretically, we assume that both aspects have a co-occurring yet distinct influence on the legitimization of content. This was supported by our exploratory results (see Table A1 in our online appendix on OSF: https://osf.io/zgqkd/)

Figure 3. *Attention to the Picture and Text Elements per Condition.*



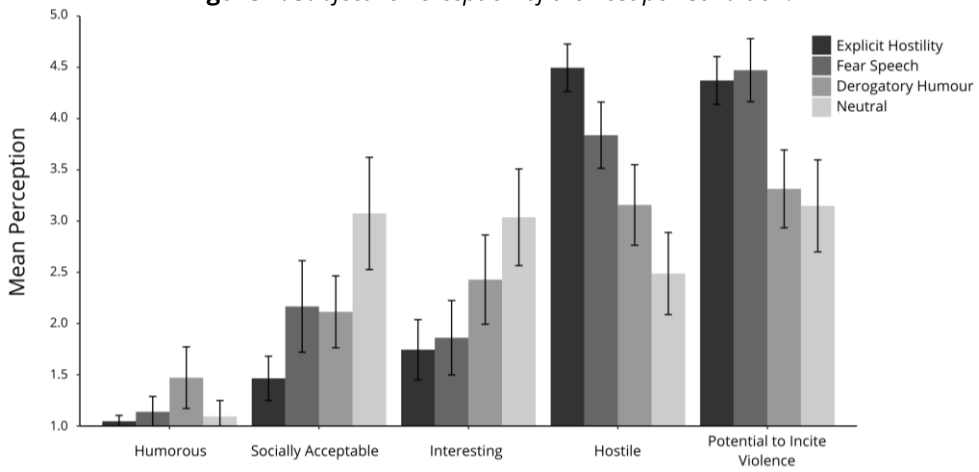
Note. Dwell time: $F(3, 137) = 2.90$, $p = .037$, $\eta^2 = .06$; Number of fixations: $F(3, 137) = 2.97$, $p = .034$, $\eta^2 = .06$. Significant differences between conditions are marked with brackets ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; family-wise error corrected, Holm).

Perception of Posts

To explore social media users' subjective perceptions of the posts containing hostility (H1), we calculated multivariate analyses of variance (MANOVA) comparing the different perceptional dimensions between the experimental conditions (explicit hostility, fear speech, derogatory humor, neutral control condition).

We found significant differences for all perceptual dimensions, including the perception as humorous ($F(3, 137) = 4.79$, $p = .003$; $\eta^2 = .09$), socially acceptable ($F(3, 137) = 11.78$, $p < .001$; $\eta^2 = .21$), interesting ($F(3, 137) = 9.08$, $p < .001$; $\eta^2 = .17$), and hostile ($F(3, 137) = 27.69$, $p < .001$; $\eta^2 = .38$), as well as for the perceived potential to incite violence ($F(3, 137) = 17.36$, $p < .001$; $\eta^2 = .28$). In line with our assumptions, both fear speech and derogatory humor were perceived significantly more socially acceptable (explicit hostility: $M = 1.47$, $SD = 0.70$; fear speech: $M = 2.17$, $SD = 1.32$; derogatory humor: $M = 2.11$; $SD = 1.02$), but less hostile than explicit hostility (explicit hostility: $M = 4.50$, $SD = 0.75$; fear speech: $M = 3.84$, $SD = 0.95$, derogatory humor: $M = 3.16$, $SD = 1.14$). Moreover, participants rated the humorous condition as significantly more humorous ($M = 1.47$; $SD = 0.87$) and interesting ($M = 2.43$, $SD = 1.27$) than the explicit hostility (humorous: $M = 1.05$, $SD = 0.18$; interesting: $M = 1.74$, $SD = 0.95$), but also more so than the fear speech condition (humorous: $M = 1.14$, $SD = 0.44$, interesting: $M = 1.86$, $SD = 1.07$). Similarly, derogatory humor was perceived to have a lower potential to incite violence against migrants than the other forms of hostility (derogatory humor: $M = 3.31$, $SD = 1.11$; fear speech: $M = 4.47$, $SD = 0.91$, explicit hostility: $M = 4.37$, $SD = 0.76$). However, the perception of fear speech and explicit hostility did not differ in this aspect (see Figure 4 and Table 1 for details).

Figure 4. *Subjective Perception of the Post per Condition.*



Note. Humorous: $F(3, 137) = 4.79$, $p = .003$; $\eta^2 = .09$; Socially acceptable: $F(3, 137) = 11.78$, $p < .001$; $\eta^2 = .21$; Interesting: $F(3, 137) = 9.08$, $p < .001$; $\eta^2 = .17$; Hostile*:* $F(3, 137) = 27.69$, $p < .001$; $\eta^2 = .38$; Potential to incite violence: $F(3, 137) = 17.36$, $p < .001$; $\eta^2 = .28$.

Table 1. *Pairwise Comparisons of Perception Variables Between Conditions.*

Condition 1	Dependent Variable	Condition 2	Mean diff.	*p*-value
Derogatory Humor	Humorous	Explicit Hostility	0.43	.003
		Fear Speech	0.33	.036
		Neutral	0.38	.029
	Socially Acceptable	Explicit Hostility	0.65	.022
		Fear Speech	−0.05	.842
		Neutral	−0.96	.005
	Interesting	Explicit Hostility	0.68	.031
		Fear Speech	0.57	.101
		Neutral	−0.61	.101
	Hostile	Explicit Hostility	−1.34	< .001
		Fear Speech	−0.68	.009
		Neutral	0.67	.009
	Potential to Incite Violence	Explicit Hostility	−1.06	< .001
		Fear Speech	−1.16	< .001
		Neutral	0.17	1.000
Fear Speech	Humorous	Explicit Hostility	0.09	1.000
		Neutral	0.05	1.000
	Socially Acceptable	Explicit Hostility	0.70	.017
		Neutral	−0.91	.006
	Interesting	Explicit Hostility	0.12	.643
		Neutral	−1.18	< .001
	Hostile	Explicit Hostility	−0.66	.009
		Neutral	1.35	< .001
	Potential to Incite Violence	Explicit Hostility	0.10	1.000
		Neutral	1.32	< .001
Explicit Hostility	Humorous	Neutral	−0.05	1.000
	Socially Acceptable	Neutral	−1.61	< .001
	Interesting	Neutral	−1.29	< .001
	Hostile	Neutral	2.01	< .001
	Potential to Incite Violence	Neutral	1.22	< .001

Note. p-values are family-wise error corrected (Holm).

Thus, we find the general tendencies hypothesized in H1, indicating that implicit forms of hostility are perceived as more justified and less harmful, though not consistently across all tested variables.

To examine the influence of cavalier humor beliefs on users' perceptions of implicit (fear speech, derogatory humor) and explicit hostility (RQ2), we computed block-wise linear regression models for the perception of the posts as hostile (Table 2), socially acceptable (Table 3), potentially violence inducing (Table 4) and humorous (Table 5). We ran the regression analyses each with two independent blocks, including (1) main effects for CHBs and the type of hostility (fear speech, derogatory humor, neutral control condition) with explicit hostility serving as the reference category and (2) the interaction of CHBs with the type of hostility. A preliminary inspection of the zero-order correlations showed no intercorrelation between the independent variables. No multicollinearity was observed (VIFs < 2).

Table 2. *Hierarchical Regression Results for Hostility Perception.*

	Direct Effects (Model 1)				Interaction (Model 2)			
Variable	*b*	β	*SE*	*p*	*b*	β	*SE*	*p*
Constant	5.39***	0.76	0.32	< .001	5.00***	0.75	0.48	< .001
CHBs[a]	−0.33**	−0.21	0.10	.002	-0.19	−0.12	0.17	.269
Fear Speech	−0.73***	−0.61	0.21	.001	0.63	−0.64	0.74	.397
Derogatory Humor	−1.37***	−1.14	0.21	< .001	−1.52*	−1.13	0.74	.041
Neutral	−2.05***	−1.70	0.23	< .001	−1.33	−1.69	0.88	.133
CHBs x Fear Speech					−0.53(*)	−0.34	0.28	.055
CHBs x Derogatory Humor					0.06	0.04	0.26	.813
CHBs x Neutral					−0.27	−0.17	0.32	.402
Observations	141				141			
R^2	.42				.44			
Adjusted R^2	.40				.41			
F Statistic	24.590*** (*df* = 4; 136)				15.009*** (*df* = 7; 133)			

Note. (x)*p* < 0.1; *p < .05; **p < .01; ***p < .001; a mean-centered.

Table 3. *Hierarchical Regression Results for Social Acceptability Perception.*

	Direct Effects (Model 1)				Interaction (Model 2)			
Variable	*b*	β	*SE*	*p*	*b*	β	*SE*	*p*
Constant	1.42***	−0.56	0.16	< .001	1.44***	−0.55	0.16	< .001
CHBs[a]	0.40**	0.25	0.12	.001	0.28	0.17	0.19	.148
Fear Speech	0.79**	0.65	0.24	.001	0.76**	0.62	0.24	.002
Derogatory Humor	0.69**	0.56	0.24	.005	0.68**	0.55	0.24	.006
Neutral	1.66***	1.35	0.26	< .001	1.65***	1.35	0.26	< .001
CHBs x Fear Speech					−0.05	−0.03	0.31	.869
CHBs x Derogatory Humor					0.03	0.02	0.30	.922
CHBs x Neutral					0.87*	0.54	0.36	.019
Observations	141				141			
R^2	.27				.30			
Adjusted R^2	.24				.26			
F Statistic	12.273*** (*df* = 4; 136)				8.186*** (*df* = 7; 133)			

Note. (x)*p* < 0.1; *p < .05; **p < .01; ***p < .001; a mean-centered.

Table 4. *Hierarchical Regression Results for Perceived Potential to Incite Violence.*

	Direct Effects (Model 1)				Interaction (Model 2)			
Variable	*b*	β	*SE*	*p*	*b*	β	*SE*	*p*
Constant	4.40***	0.44	0.15	< .001	4.40***	0.45	0.15	< .001
CHBs[a]	−0.24*	−0.16	0.11	.025	−0.28	−0.19	0.17	.110
Fear Speech	0.04	0.04	0.22	.840	0.02	0.01	0.22	.945
Derogatory Humor	−1.08***	−0.96	0.22	< .001	−1.09***	−0.97	0.22	< .001
Neutral	−1.25***	−1.11	0.23	< .001	−1.25***	−1.12	0.24	< .001
CHBs x Fear Speech					−0.15	−0.10	0.29	.589
CHBs x Derogatory Humor					0.18	0.12	0.27	.522
CHBs x Neutral					0.20	0.13	0.33	.554
Observations	141				141			
R^2	.30				.31			
Adjusted R^2	.28				.27			
F Statistic	14.694*** (*df* = 4; 136)				8.521*** (*df* = 7; 133)			

Note. (x)*p* < 0.1; *p < .05; **p < .01; ***p < .001; a mean-centered.

Table 5. *Hierarchical Regression Results for Humor Perception.*

	Direct Effects (Model 1)				Interaction (Model 2)			
Variable	*b*	β	*SE*	*p*	*b*	β	*SE*	*p*
Constant	1.03***	−0.27	0.08	< .001	1.04***	−0.26	0.08	< .001
CHBs[a]	0.12*	0.16	0.06	.044	0.02	0.03	0.09	.801
Fear Speech	0.12	0.22	0.12	.314	0.10	0.19	0.12	.383
Derogatory Humor	0.44***	0.79	0.12	< .001	0.42***	0.77	0.12	< .001
Neutral	0.06	0.11	0.13	.641	0.05	0.09	0.13	.704
CHBs x Fear Speech					0.03	0.05	0.15	.824
CHBs x Derogatory Humor					0.39**	0.54	0.15	.009
CHBs x Neutral					−0.08	−0.12	0.18	.636
Observations	141				141			
R^2	.12				.18			
Adjusted R^2	.10				.14			
F Statistic	4.703** (*df* = 4; 136)				4.179*** (*df* = 7; 133)			

Note. (x)*p* < 0.1; *p < .05; **p < .01; ***p < .001; a mean-centered.

Examining the models from Block 1 (without interaction effects) reveals the same tendencies in how the experimental groups influence post perception as previously discussed. Moreover, higher levels of cavalier humor beliefs (CHBs) are shown to be positively associated with perceptions of social acceptability (b = 0.40, p = .001; Table 3) and humor (b = 0.12, p = .044; Table 5) while being negatively associated with perceptions of hostility (b = −0.33, p = .002; Table 2) and the potential to incite violence (b = −0.24, p = .025; Table 4). Considering CHBs as moderator in the experimental groups in Block 2 of the models reveals that higher levels of CHBs are associated with higher perceptions of humor only regarding derogatory humor compared to explicit hostility (b = 0.39, p = .009; Table 5). Moreover, higher levels of CHBs are associated with higher perceptions of social acceptability regarding the neutral control conditions compared to explicit hostility (b = 0.87, p = .019; Table 3). Due to the small sample and effect size, however, these interaction results should be interpreted with caution. Thus, answering RQ2, we did not find any effects of CHBs on the perception of the different stimuli, except for the most evident ones concerning derogatory humor. No associations between CHBs and perceptions of fear speech were observed in this study.

In an additional exploratory analysis, we tested whether CHBs were related to participants' visual attention. The results indicated no substantial relationships between CHBs and the visual attention measures (see Table A2–A4 in our online appendix on OSF: https://osf.io/zgqkd/).

Discussion

Summary of Findings and Implications

This study explored the potential of two forms of implicit communication in legitimizing hostility on social media platforms: fear speech and derogatory humor. We argued that while these forms do not explicitly express hatred against outgroups, they nonetheless promote hostile ideologies by communicating emotions that resonate with a broader audience and, as exemplified in the JSM (Crandall & Eshleman, 2003), are able to justify the expression of prejudices. Our theoretical framework further draws on legitimization strategies previously analyzed through critical discourse analyses in various areas of (political) communication (Reyes, 2011; van Leeuwen, 2007; van Leeuwen & Wodak, 1999), including social media posts (Davis et al., 2016; Koivukoski et al., 2025; Recuero & Soares, 2022; Ross & Rivers, 2017). Building on literature about outgroup delegitimization in intergroup contexts (Bar-Tal & Hammack, 2012; Hodson & MacInnis, 2016), we extended this perspective to examine how these strategies influence social media audiences in their reception of hostile content. Specifically, we focused on (1) social media users' visual attention to and (2) their perception of such content, arguing that the initial moment of reception is critical in increasing susceptibility and, thus, legitimization.

The findings of our study indicate that implicit forms of hostility are highly effective in legitimizing hostile content and facilitating the normalization of underlying extreme views. By using derogatory humor or fear speech rather than explicit forms of hostility, communicators' strategic communication can effectively downplay the severity of their message. Our findings suggest that such differences emerge in the reception situation when audiences encounter hostile messages. This is very well in line with the predictions of the JSM, which posits that individuals are more likely to express prejudices when they find a justification for doing so (Crandall & Eshleman, 2003).

Regarding the initial moment of encountering hostile content—recipient's visual attention to the content—eye-tracking measurements showed that implicit forms of hostility received at least as much, and in the case of derogatory humor, even more attention, measured against the text as the central semantic element of the social media posts. It must be noted that humorous content may generally require more cognitive processing capacity (Suls, 1972), potentially reflected in heightened visual attention here.

Regarding the second examined step—recipient's perception of the content—derogatory humor tended to be perceived as a less serious issue overall, which is consistent with previous research on the perception of humorous expressions of hatred (Schmid, 2025; Schmid & Greipl, 2025; Yeon & Lee, 2021). Thus, even if heightened attention results from complex processing, our data suggests it led to acceptance rather than rejection. This highlights the importance of the simultaneity of the two explored aspects: humorous content both attracts attention and is perceived as less harmful. The general pattern of participants' perceptions aligns with what we had expected: a descending trend, ranging from explicit to fear speech, then to derogatory humor, and finally the neutral social media posts, with significant differences between each condition. Together with the higher perception of derogatory humor as socially acceptable, this finding highlights the potential of humor in legitimizing hostility, and, correspondingly, to delegitimizing social outgroups (Hodson & MacInnis, 2016).

The strategic potential of fear speech becomes especially evident when considering its perceived capacity to incite violence. While derogatory humor was rated similarly to neutral posts in this regard, fear speech was viewed as having the greatest potential to provoke violence against migrants that were targeted in the post. Particularly notable is the interplay of the perceptional dimensions, especially regarding the social acceptability of fear speech. Despite its acknowledged potential for inducing violence, fear speech appears to be met with a tolerance comparable to that of derogatory humor. This paradox underscores the potential of fear speech in legitimizing hostility to the point where even acts of violence might seem justifiable (Buyse, 2014; Sayimer & Derman, 2017). In this way, fear speech operates as a form of symbolic violence, subtly reinforcing systems of exclusion and (violent) oppression. From this perspective, in addition to observing perpetrators of physical violence, monitoring propagandists of symbolic violence by using delegitimizing narratives is crucial as these narratives may ultimately lead to physical violence (Bar-Tal & Hammack, 2012). It is important to note, however, that the higher perceived potential for violence in the fear speech condition does not necessarily imply that such content will in fact incite

more violent behavior than derogatory humor. Our measure captures recipients' subjective assessments of the posts' dangerousness rather than actual behavioral outcomes. While these perceptions are meaningful in themselves, since they may influence tolerance, intervention, or policy responses, future research should examine behavioral consequences more directly, for example through measures of intervention willingness, behavioral intentions, or longitudinal designs. From a societal perspective, this highlights that both fear speech and derogatory humor may be consequential, albeit through different pathways: one by amplifying threat perceptions, the other by trivializing hostility.

Finally, CHBs as an attitudinal disposition indicated a rather uniform effect on the perceptual level. As expected, cavalier humor beliefs, reflecting a tendency toward social dominance, were associated with reduced perceptions of hostility and perceived potential for inciting violence, while increasing perceived social acceptability and humor appreciation. Notably, however, these effects largely did not reach statistical significance when cavalier humor beliefs are considered a moderator within the experimental groups. Some tendencies were observed: for instance, participants exposed to derogatory humor found the material funnier depending on their degree of cavalier humor beliefs. Thus, in line with previous research (Hodson & MacInnis, 2016), CHBs specifically affect recipients' perception of borderline humor.

From a long-term perspective, symbolic violence on social media platforms becomes self-perpetuating. Implicit forms of hostility are not only more readily accepted by recipients but are also more difficult to detect (ElSherief et al., 2021). Due to their indirect and subtle packaging, such symbolic acts of violence are often categorized as borderline content (Gillett et al., 2022) rather than explicitly harmful, rendering them less susceptible to platform regulation. Consequently, this type of content continues to gain visibility and reach, further reinforcing the perceived legitimacy of toxic and violent online discourses (Recuero, 2024). This chain of events can have significant consequences. Beyond shaping users' perception, toxic online discourses may also encourage users to adopt similar hostile commenting behaviors (Kim et al., 2021). Ford and Ferguson's (2004) prejudiced norm theory suggests that encountering derogatory humor not only increases the tolerance of discrimination against the targeted group and reinforces negative stereotypes but also heightens individuals' willingness to discriminate oneself as well as the propensity to commit violence. In the broader context of mainstreaming processes, these dynamics could lead to a gradual deterioration of online discourse, ultimately normalizing explicit forms of hostility and making them appear as legitimate expressions within mainstream conversation (Rothut et al., 2024). This highlights the need for platform policies and moderation practices to account for the indirect nature of symbolic violence, ensuring that content is evaluated not only for overt harm but also for its potential to normalize and perpetuate hostility through nuanced, subtle shifts.

Limitations and Directions to Future Research

A limitation of the present study is that we did not directly assess recipients' emotional responses (e.g., fear, anxiety, or amusement). While our perceptual measures capture downstream legitimization processes theorized to be shaped by emotional communication, future research should explicitly integrate affective indicators such as self-reported or psychophysiological measures to more precisely disentangle the role of emotions in legitimization dynamics. Moreover, the study did not include a direct manipulation check of the stimuli to verify whether participants recognized the humor- or fear-based cues within the stimuli. Additionally, the findings are restricted by our relatively small and homogeneous sample of 141 participants, all recruited in a university setting. This introduces potential bias, as the participants were highly educated. This led to low variability in their perception of hostile posts and may also account for some inconsistencies regarding CHBs. While the study was sufficiently powered to detect small-to-medium main effects, its power to detect more subtle interaction effects—particularly involving individual dispositions such as cavalier humor beliefs—was constrained. Future research with larger and more heterogeneous samples will be needed to more comprehensively test these interactive processes. A larger-scale study could also include additional stimuli and target groups for investigation, allowing to evaluate if different topics attract varying degrees of attention and, by using more than two stimulus versions per condition, better isolate the effects of the speech type from stimulus-specific effects. However, measuring participants' visual attention using eye-tracking introduces its own limitations. Since the participants of our study were in a laboratory setting and aware that their eye movements were being tracked, the results may not fully translate to real-world social media use. This issue is compounded by the fact that we used a fictitious social media platform with unfamiliar users, overlooking the influence of knowing the communicator personally or recognizing his or her

authority (e.g., as a politician), which highly contributes the legitimization of hostile discourses as well (Recuero, 2024).

Conclusion

Using a combination of eye-tracking and survey data, this study demonstrates how fear speech and derogatory humor can subtly legitimize hostility on social media. Drawing on extensive literature in critical discourse analysis and psychological research on prejudice and intergroup conflict, our findings suggest that these forms of symbolic violence resonate broadly and are more socially accepted and justified than explicit hostile content. Given the cumulative impact of implicit hostility, our study underscores the need for monitoring of such content, which circulates with relative ease. Although it may not always constitute a direct legal or policy violation, this content plays a significant role in normalizing exclusionary ideologies—a process further amplified by the affordances of social media platforms.

Footnote

[1] To assess equivalence of the stimulus sets (two stimulus motives), we conducted 2 (stimulus set) × 4 (experimental condition) ANOVAs for all dependent variables. No significant main effects of stimulus sets were found (all Fs < 1.35, ps > .24), indicating that the two sets did not differ inherently in their baseline perceptions. For hostility perception, perceived potential to incite violence, and humor perception, no interactions between set and experimental condition were observed (ps > .24). A significant interaction emerged for social acceptability, $F(3, 133) = 2.99$, $p = .033$, but the primary effect of experimental condition remained highly robust. The significant interaction for social acceptability was driven by minor variations in specific stimulus pairings but did not alter the direction of the main effects. Crucially, the experimental manipulation explained more than three times the variance (21.6%) compared to the interaction (6.3%), supporting the overall generalizability of the findings across stimulus sets. Analyses of gaze behavior confirmed that the stimulus sets were visually balanced. Across all Areas of Interest (AoIs), there were no significant main effects of stimulus set on dwell time or number of fixations (all ps > .28, $\eta_p^2 < .009$), indicating that attentional allocation was not biased by the specific stimulus sets. While minor interactions between stimulus set and experimental condition emerged for the Picture AoI, the absence of main effects for stimulus set ensures that the primary findings are not attributable to inherent differences in visual complexity.

Conflict of Interest

The authors have no conflicts of interest to declare.

Use of AI Services

The authors declare they have used AI services for grammar correction and minor style refinements. They carefully reviewed all suggestions from these services to ensure the original meaning and factual accuracy were preserved.

Authors' Contribution

Ursula Kristin Schmid: conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft. **Simon Greipl:** conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing—review & editing. **Diana Rieger:** conceptualization, project administration, resources, supervision, writing—review & editing.

Acknowledgement

The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement

The data underlying this article will be shared on reasonable request to the corresponding author.

References

Albertson, B., & Gadarian, S. K. (2015). *Anxious politics: Democratic citizenship in a threatening world*. Cambridge University Press. https://doi.org/10.1017/CBO9781139963107

Bar-Tal, D. (2013). *Intractable conflicts: Socio-psychological foundations and dynamics.* Cambridge University Press. https://doi.org/10.1017/CBO9781139025195

Bar-Tal, D., & Hammack, P. L. (2012). Conflict, delegitimization, and violence. In L. R. Tropp (Ed.), *The Oxford handbook of intergroup conflict* (pp. 29–52). https://doi.org/10.1093/oxfordhb/9780199747672.013.0003

Ben-David, A., & Matamoros-Fernandez, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193. https://ijoc.org/index.php/ijoc/article/view/3697/1585

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition: When authoritarians confront prejudice. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Billig, M. (2001). Humour and hatred: The racist jokes of the Ku Klux Klan. *Discourse & Society*, *12*(3), 267–289. https://doi.org/10.1177/0957926501012003001

Borah, P. (2016). Political Facebook use: Campaign strategies used in 2008 and 2012 presidential elections. *Journal of Information Technology & Politics*, *13*(4), 326–338. https://doi.org/10.1080/19331681.2016.1163519

Bourdieu, P. (1991). *Language and symbolic power.* Polity Press.

Buie, H. S., Ford, T. E., Olah, A. R., Argüello, C., & Mendiburo-Seguel, A. (2022). Where's your sense of humor? Political identity moderates evaluations of disparagement humor. *Group Processes & Intergroup Relations*, *25*(5), 1395–1411. https://doi.org/10.1177/1368430221998792

Buyse, A. (2014). Words of violence: "Fear speech," or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, *36*(4), 779–797. https://doi.org/10.1353/hrq.2014.0064

Chovanec, J. (2021). 'Re-educating the Roma? You must be joking. . .': Racism and prejudice in online discussion forums. *Discourse & Society*, *32*(2), 156–174. https://doi.org/10.1177/0957926520970384

Coronel, J. C., O'Donnell, M. B., Pandey, P., Delli Carpini, M. X., & Falk, E. B. (2021). Political humor, sharing, and remembering: Insights from neuroimaging. *Journal of Communication*, *71*(1), 129–161. https://doi.org/10.1093/joc/jqaa041

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Crilley, R., & Chatterje-Doody, P. N. (2021). From Russia with lols: Humour, RT, and the legitimation of Russian foreign policy. *Global Society*, *35*(2), 269–288. https://doi.org/10.1080/13600826.2020.1839387

Davies, H., & Ilott, S. (2018). *Comedy and the politics of representation: Mocking the weak*. Springer International Publishing. https://doi.org/10.1007/978-3-319-90506-8

Davis, C. B., Glantz, M., & Novak, D. R. (2016). "You can't run your SUV on cute. Let's go!": Internet memes as delegitimizing discourse. *Environmental Communication*, *10*(1), 62–83. https://doi.org/10.1080/17524032.2014.991411

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345–363). https://doi.org/10.18653/v1/2021.emnlp-main.29

Feischmidt, M., & Hervik, P. (2015). Mainstreaming the extreme: Intersecting challenges from the far right in Europe. *Intersections. East European Journal of Society and Politics*, *1*(1). https://doi.org/10.17356/ieejsp.v1i1.80

Fielitz, M., Donner, C., Bitzmann, H., Brodersen, W., Marcks, & Sick, H. (2024, January 29). Five shades of hate: Gruppenbezogene Abwertung in Zeiten der Memifizierung [Five shades of hate: Group-Based devaluation in the age of memification]. *Machine Against the Rage*. https://www.doi.org/10.58668/matr/05.2

Ford, T. E., & Ferguson, M. (2004). Social consequences of disparagement humor: A prejudiced norm theory. *Personality and Social Psychology Review*, *8*(1), 79–94. https://doi.org/10.1207/S15327957PSPR0801_4

Ford, T. E., Boxer, C. F., Armstrong, J., & Edel, J. R. (2008). More than "just a joke": The prejudice-releasing function of sexist humor. *Personality and Social Psychology Bulletin*, *34*(2), 159–170. https://doi.org/10.1177/0146167207310022

Freiheit, M., & Zick, A. (2022). Die Rolle von islamistischen Gruppen und Milieus in der Hinwendung und Radikalisierung von jungen Menschen [The role of Islamist groups and milieus in the recruitment and radicalization of young people]. In B. Milbradt, A. Frank, F. Greuel, & M. Herding (Eds.), *Handbuch Radikalisierung im Jugendalter. Phänomene, Herausforderungen, Prävention [Handbook radicalization among youth: phenomena, challenges, and prevention]* (pp. 247–262). Verlag Barbara Budrich.

Gillett, R., Stardust, Z., & Burgess, J. (2022). Safety for whom? Investigating how platforms frame and perform safety and harm interventions. *Social Media + Society*, *8*(4), 1–12. https://doi.org/10.1177/20563051221144315

Greipl, S., Hohner, J., Schulze, H., Schwabl, P., & Rieger, D. (2024). "You are doomed!" Crisis-specific and dynamic use of fear speech in protest and extremist radical social movements. *Journal of Quantitative Description: Digital Media*, *4*, 1–46. https://doi.org/10.51685/jqd.2024.icwsm.8

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Hodson, G., & MacInnis, C. C. (2016). Derogating humor as a delegitimization strategy in intergroup contexts. *Translational Issues in Psychological Science*, *2*(1), 63–74. https://doi.org/10.1037/tps0000052

Hodson, G., Rush, J., & Macinnis, C. (2010). A joke is just a joke (except when it isn't): Cavalier humor beliefs facilitate the expression of group dominance motives. *Journal of Personality and Social Psychology*, *99*(4), 660–682. https://doi.org/10.1037/a0019627

Kakavand, A. E. (2024). Far-right social media communication in the light of technology affordances: A systematic literature review. *Annals of the International Communication Association*, *48*(1), 37–56. https://doi.org/10.1080/23808985.2023.2280824

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

King, A. J., Bol, N., Cummins, R. G., & John, K. K. (2019). Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Communication Methods and Measures*, *13*(3), 149–177. https://doi.org/10.1080/19312458.2018.1558194

Klein, A. (2021). Social networks and the challenge of hate disguised as fear and politics. *Journal for Deradicalization*, *26*, 1–33. https://journals.sfu.ca/jd/index.php/jd/article/view/431/259

Knobloch-Westerwick, S. (2014). *Choice and preference in media use: Advances in selective exposure theory and research*. Routledge. https://doi.org/10.4324/9781315771359

Koivukoski, J., Laaksonen, S.-M., Zareff, J., & Knuutila, A. (2025). Challenging and enhancing legitimacy through humor: A comparative study of candidates' use of humor on Facebook before the 2019 and 2023 Finnish parliamentary elections. *Alternatives: Global, Local, Political, 50*(1), 74–93. https://doi.org/10.1177/03043754241272272

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, *50*(1), 46–70. https://doi.org/10.1111/j.1460-2466.2000.tb02833.x

Lintott, S. (2016). Superiority in humor theory. *The Journal of Aesthetics and Art Criticism*, *74*(4), 347–358. https://doi.org/10.1111/jaac.12321

Marcks, H., & Pawelz, J. (2022). From myths of victimhood to fantasies of violence: How far-right narratives of imperilment work. *Terrorism and Political Violence*, *34*(7), 1415–1432. https://doi.org/10.1080/09546553.2020.1788544

Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgment*. University of Chicago Press.

Matamoros-Fernández, A., Bartolo, L., & Troynar, L. (2023). Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1677

McSwiney, J., & Sengul, K. (2024). Humor, ridicule, and the far right: Mainstreaming exclusion through online animation. *Television & New Media*, *25*(4), 315–333. https://doi.org/10.1177/15274764231213816

Meiering, D., Dziri, A., Foroutan, N., Teune, S., Lehnert, E., & Abou Taam, M. (2018). *Brückennarrative—Verbindende Elemente für die Radikalisierung von Gruppen* [Bridging narratives -connecting factors in group radicalization]. Hessische Stiftung Friedens- und Konfliktforschung. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59476-6

Mudde, C. (2019). *The far right today*. Polity.

Nann, L., Udupa, S., & Wisiorek, A. (2024). Online anti-immigrant discourse in Germany: Ethnographically backed analysis of user comments. *Frontiers in Communication*, *9,* Article 135502. https://doi.org/10.3389/fcomm.2024.1355025

Oddo, J. (2011). War legitimation discourse: Representing 'Us' and 'Them' in four US presidential addresses. *Discourse & Society*, *22*(3), 287–314. https://doi.org/10.1177/0957926510395442

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199999736.001.0001

Pérez, R. (2013). Learning to make racism funny in the 'color-blind' era: Stand-up comedy students, performance strategies, and the (re)production of racist jokes in public. *Discourse & Society*, *24*(4), 478–503. https://doi.org/10.1177/0957926513482066

Prusaczyk, E., & Hodson, G. (2020). "To the moon, Alice": Cavalier humor beliefs and women's reactions to aggressive and belittling sexist jokes. *Journal of Experimental Social Psychology*, *88*, Article 103973. https://doi.org/10.1016/j.jesp.2020.103973

Recuero, R. (2015). Social media and symbolic violence. *Social Media + Society*, *1*(1). https://doi.org/10.1177/2056305115580332

Recuero, R. (2024). The platformization of violence: Toward a concept of discursive toxicity on social media. *Social Media + Society*, *10*(1), 1–9. https://doi.org/10.1177/20563051231224264

Recuero, R., & Soares, F. B. (2022). #VACHINA: How politicians help to spread disinformation about COVID-19 vaccines. *Journal of Digital Social Research*, *4*(1), 73–97. https://doi.org/10.33621/jdsr.v4i1.112

Reyes, A. (2011). Strategies of legitimization in political discourse: From words to actions. *Discourse & Society*, *22*(6), 781–807. https://doi.org/10.1177/0957926511419927

Rieger, D., Frischlich, L., & Bente, G. (2013). *Propaganda 2.0: Psychological effects of right-wing and Islamic extremist internet videos*. Polizei + Forschung.

Ross, A. S., & Rivers, D. J. (2017). Digital cultures of political participation: Internet memes and the discursive delegitimization of the 2016 U.S Presidential candidates. *Discourse, Context & Media*, *16*, 1–11. https://doi.org/10.1016/j.dcm.2017.01.001

Rothut, S., Schulze, H., Rieger, D., Lechner, M., & Naderer, B. (2026). Protest movements and the mainstreaming of radical and extremist ideologies: The case of COVID-19 protests. *Information, Communication & Society.* Advance online publication. https://doi.org/10.1080/1369118X.2026.2654668

Rothut, S., Schulze, H., Rieger, D., & Naderer, B. (2024). Mainstreaming as a meta-process: A systematic review and conceptual model of factors contributing to the mainstreaming of radical and extremist positions. *Communication Theory*, *34*(2), 49–59. https://doi.org/10.1093/ct/qtae001

Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021). "Short is the road that leads from fear to hate": Fear speech in Indian WhatsApp groups. In *Proceedings of the Web Conference 2021, WWW '21* (pp. 1110–1121). https://doi.org/10.1145/3442381.3450137

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

Santana, A. D., & Hopp, T. (2022). Seeing red: Reading uncivil news comments guided by personality characteristics. *Newspaper Research Journal*, *43*(2), 196–216. https://doi.org/10.1177/07395329221094662

Sayimer, İ., & Derman, M. R. (2017). Syrian refugees as victims of fear and danger discourse in social media: A Youtube analysis. *Global Media Journal TR Edition*, *8*(15), 384–403.

Schmid, U. K. (2025). Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society, 27*(3), 1588–1606. https://doi.org/10.1177/14614448231198169

Schmid, U. K., & Greipl, S. (2025). The thin line between harmful and benign: Perceptions of humorous and non-humorous hate speech in humorous and neutral social media contexts. *Media Psychology*, 1–26. Advance online publication. https://doi.org/10.1080/15213269.2025.2580291

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2024). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, *26*(5), 2614–2632. https://doi.org/10.1177/14614448221091185

Schmid, U. K., Schulze, H., & Drexel, A. (2025). Memes, humor, and the far right's strategic mainstreaming. *Information, Communication & Society, 28*(4), 537–556. https://doi.org/10.1080/1369118x.2024.2329610

Schmitt, J. B., Harles, D., & Rieger, D. (2020). Themen, motive und mainstreaming in rechtsextremen online-memes [Themes, motifs, and mainstreaming in far-right online memes]. *M&K Medien & Kommunikationswissenschaft*, *68*(1–2), 73–93. https://doi.org/10.5771/1615-634X-2020-1-2-73

Schulze, H., Greipl, S., Hohner, J., & Rieger, D. (2024). Social media and radicalization: An affordance approach for cross-platform comparison. *M&K Medien & Kommunikationswissenschaft*, *72*(2), 187–212. https://doi.org/10.5771/1615-634X-2024-2-187

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression* (pp. x–403). Cambridge University Press. https://doi.org/10.1017/CBO9781139175043

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 81–100). Academic Press.

van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse & Communication*, *1*(1), 91–112. https://doi.org/10.1177/1750481307071986

van Leeuwen, T., & Wodak, R. (1999). Legitimizing immigration control: A discourse-historical analysis. *Discourse Studies*, *1*(1), 83–118. https://doi.org/10.1177/1461445699001001005

Weaver, S. (2011). *The rhetoric of racist humour: US, UK and global race joking*. Routledge.

Wodak, R. (2021). *The politics of fear: The shameless normalization of far-right discourse*. SAGE Publications Ltd. https://doi.org/10.4135/9781529739664

Yeon, J., & Lee, H. (2021). When hate meets humor: The effect of humor to amplify hatred and disgust toward outgroup and the implications for gender conflicts in South Korea. *The Journal of Inequality and Democracy*, *4*(1), 20–47. https://doi.org/10.18854/kpsr.2020.54.4.009

Žižek, S. (2007). *Violence: Six sideways reflections*. Picador.

About Authors

Ursula Kristin Schmid, MA, is a scientific researcher at the Department of Media and Communication at LMU Munich. Her research focuses on characteristics and perceptions of online hate speech and counter speech. In her PhD thesis, she researches the perception of hate speech disguised as humor.

https://orcid.org/0000-0002-1892-002X

Simon Greipl, M.Sc., is a scientific researcher at the Department of Media and Communication at LMU Munich. His research interests include the indication of radicalization dynamics in online environments, especially in relation to gaming and its communities.

https://orcid.org/0000-0002-5652-8889

Diana Rieger is a Professor at the Department of Media and Communication at LMU Munich. In her research, she investigates indicators for online radicalization dynamics, such as hate speech, as well as the impact of toxic online communication on internet users.

https://orcid.org/0000-0002-2417-0480

Correspondence to
Ursula Kristin Schmid, Department of Media and Communication, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany, ursula.schmid@ifkw.lmu.de