

With a Little Help From AI?! Governing Visual Hate Speech Through AI-Assisted Counter-Image Practices

Stefano Pedrazzi ¹  and Franziska Oehmer-Pedrazzi ^{2,3} 

¹ Department of Communication and Media Research, University of Fribourg, Switzerland

² MILEVA Institut für Digitales und Gesellschaft, Switzerland

³ University of Applied Sciences of the Grisons, Switzerland

Correspondence: Stefano Pedrazzi (stefano.pedrazzi@unifr.ch)

Submitted: 31 March 2026 **Accepted:** 4 August 2026 **Published:** 21 October 2026

Issue: This article is part of the issue “The Role of AI for Counter Speech: Detection, Intervention, and Risks” edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at <https://doi.org/10.17645/mac.i500>

Abstract

The attention-grabbing and memorable nature of visuals makes them a powerful vehicle for spreading hate online. As platforms increasingly reconsider or scale back content moderation efforts, user-centered interventions are gaining importance as complementary governance strategies. While prior research demonstrates the effectiveness of empathy-based textual counter speech, little is known about the role of visual counter speech in mitigating hateful imagery. This preregistered study examines the effectiveness of user-centered interventions against visual hate speech. Using an experimental design ($N = 666$), we tested four interventions against a no-intervention control condition in response to hateful images targeting transgender persons: (a) empathy-based textual counter speech, (b) AI-generated generic counter-images, (c) AI-generated specific counter-images replicating elements of the original, and (d) platform-issued warning labels. Results show substantial variation across interventions. Empathy-based textual counter speech and AI-generated specific counter-images showed the most consistent reductions in amplification-oriented behaviors, particularly forwarding. Visual interventions did not outperform textual counter speech, suggesting that effectiveness depends less on modality and more on the specificity and contextual alignment of the intervention. Furthermore, effectiveness varied with the intensity of the hate content. Interventions were most effective for humor-based, more ambiguously harmful content, whereas their impact diminished for more severe forms of hate. The consistent performance of AI-generated specific counter-images further points to the potential of generative AI to provide users with a rapid, low-cost, and accessible means of producing context-sensitive counter speech.

Keywords

AI; counter image; counter speech; governance; hate image; hate speech

1. Introduction

In February 2026, US President Donald Trump shared an AI-generated video on social media depicting former President Barack Obama and his wife Michelle Obama in a dehumanizing manner, triggering widespread public criticism (Shabad & Haake, 2026). This incident exemplifies three interrelated developments that are central to contemporary digital communication: First, visual communication has become a dominant form of expression in online environments. Images, memes, and short visual messages circulate widely across social media platforms and often generate higher levels of engagement than purely textual content (Crawford et al., 2021; Marquart, 2023; Schmid et al., 2024). This prominence is closely linked to both the affordances of digital platforms and the inherent characteristics of visual messages: images attract attention, are typically easy to process, and are more likely to be remembered than text-based information (Carney & Levin, 2002; Knobloch et al., 2003). As a result, visual formats have become central vehicles of communication in online discourse (Hornuff, 2020).

Second, the communicative power of visuals is not limited to informative or entertaining content. Visuals can also be used to disseminate hateful messages, defame individuals, or exclude social groups. Visual hate speech—including hostile photographs, caricatures, memes, graphics, and other forms of imagery—extends such communication beyond exclusively verbal expression and can communicate discrimination, dehumanization, or even calls for violence against individuals or groups based on characteristics such as gender, sexual orientation, ethnicity, nationality, or religion (Bin Aziz & Briel, 2024; Castaño-Pulgarín et al., 2021; Oehmer-Pedrazzi & Pedrazzi, 2024a). A growing body of research demonstrates that such hate messages can have severe consequences, ranging from psychological harm to broader societal effects such as the normalization of intolerance and the erosion of civil discourse (Arcila Calderón et al., 2024; Stahel & Baier, 2023; Unger & Unger-Sirsch, 2023).

Third, recent advances in generative AI (GenAI) have fundamentally expanded the production and circulation of visual content. Image generation systems enable users to create visually sophisticated content quickly, with comparatively little technical input and at low cost, thereby lowering barriers to participation in visual communication (Zhang et al., 2025). AI-created visuals may be used to generate and disseminate harmful content.

Against this background, the challenge of addressing harmful online communication has increasingly been framed as a problem of platform governance (Schünemann & Steiger, 2023). Rather than relying solely on state regulation, governance approaches emphasize the interaction of multiple actors—including governments, digital platforms, civil society organizations, and users—in regulating harmful content (Mayntz, 2004). Within this perspective, users themselves can play a crucial role in countering harmful communication, particularly in light of the tendency of many platforms to scale back content moderation efforts. One widely discussed user-centered strategy is counter speech, defined as a communicative response that challenges, condemns, or reframes hateful messages, thus seeking to influence how they are interpreted without suppressing them (Gennaro et al., 2025; Hangartner et al., 2021).

However, existing research on counter speech has focused predominantly on textual interventions (Gennaro et al., 2025; Hangartner et al., 2021; Wu et al., 2025), while the potential of visual counter speech remains largely underexplored. This constitutes a significant gap, given that much of contemporary online

communication—including hate speech—relies heavily on visual formats (Bin Aziz & Briel, 2024; Schmid et al., 2024). This raises the question of whether countermeasures that operate within the same visual modality may be particularly suitable for responding to visual hate speech.

Importantly, GenAI not only enables the production of harmful visual content but also opens up new possibilities for countering it. AI-based image generation tools may facilitate scalable forms of visual counter speech that directly engage with hateful imagery—for example, by reproducing and reframing visual elements of the original message or by presenting alternative, supportive narratives. Despite this potential, empirical evidence on the effectiveness of such interventions remains scarce.

To address this research gap, the present study examines whether and how different reactive governance interventions can shape users' intended responses to visual hate speech. Using a factorial online experiment with participants from the German-speaking part of Switzerland ($N = 666$), we compare four interventions—empathy-based textual counter speech, an AI-generated generic counter image, AI-generated or AI-assisted specific counter-images, and a platform-issued warning label—with a no-intervention control condition. Effectiveness is assessed in terms of users' intended interactions with hateful images, including amplification-oriented behaviors (liking, sharing, external forwarding, and supportive commenting) and two distinct forms of opposition: critical commenting as a communicative response and reporting as a platform-directed response.

By systematically comparing textual, visual, and platform-based interventions, this study makes three contributions. First, it advances research on online hate speech by focusing specifically on visual hate images, a form of harmful communication that has received comparatively limited empirical attention. Second, it contributes to scholarship on platform governance by comparing the effectiveness of platform-based and user-centered interventions within a common experimental framework. Third, it provides empirical evidence on the effectiveness of AI-generated or AI-assisted counter-images as a scalable tool for visual counter speech in digital environments.

2. Visual Hate Speech and its Governance

We refer to hate speech as communication that degrades, threatens, or expresses hostility toward individuals or social groups based on group-defining characteristics and may have destructive consequences (Hietanen & Eddebo, 2023). This conceptualization thus includes not only unlawful threats or incitement to violence but also discriminatory and dehumanizing messages below legal thresholds.

Hate speech may be expressed verbally, visually, or through combinations of different modes. Visual hate speech refers to hateful meanings communicated partly or entirely through visual elements, including photographs, drawings, caricatures, memes, manipulated images, and videos (Bin Aziz & Briel, 2024; Oehmer-Pedrazzi & Pedrazzi, 2024a). We use the term “hate image” for an individual visual or multimodal message that targets a person or group based on characteristics such as gender identity, sexual orientation, ethnicity, nationality, or religion. Visual hate speech is thus not simply textual hate speech accompanied by an image: the visual elements themselves may convey stereotypes, dehumanization, exclusion, or threats, while the hateful meaning may also emerge from the interplay of text and image (Oehmer-Pedrazzi & Pedrazzi, 2024a).

The interpretation of visual hate speech depends on its directness, presentation form, and communicative context (Bin Aziz & Briel, 2024; Schmid et al., 2024). Direct forms may use explicit insults, dehumanizing depictions, or calls for violence, whereas indirect forms can rely on stereotypes, symbols, irony, or shared cultural references. Humorous image–text combinations and memes can be particularly ambiguous because humor cues may make hateful content appear less hostile or allow it to be dismissed as a joke (Schmid, 2025; Schmid et al., 2024). These characteristics make visual hate speech a distinct challenge for both users and platform governance.

This study adopts a governance perspective that conceptualizes the regulation of online hate speech as a multi-actor, multi-level process. Rather than attributing responsibility solely to state institutions, governance refers to the totality of coexisting forms of collective regulation of societal affairs (Mayntz, 2004, p. 66). Applied to the context of online hate speech, this perspective encompasses the institutional, legal, technical, and social mechanisms through which harmful communication is regulated and mitigated. It explicitly acknowledges the interplay between governments, digital platforms, civil society organizations, and users as co-regulators of online communication (Puppis, 2010).

Within this framework, governance measures can be differentiated along two key dimensions: the actors involved and the temporal stage of intervention. First, multiple actors contribute to the governance of hate speech. Governments establish legal frameworks that define and sanction unlawful content, such as discriminatory or violence-inciting expressions. However, due to the transnational nature of digital communication, state regulation remains territorially bounded and often requires balancing the protection from discrimination with the constitutional guarantee of freedom of expression (Schünemann & Steiger, 2023). Platforms, in turn, operationalize governance through technical and organizational measures, including content moderation, account sanctions, algorithmic curation, and reporting mechanisms (Boberg et al., 2018). Civil society actors complement these efforts by promoting media literacy and fostering awareness of the societal consequences of hate speech. Finally, users themselves represent a crucial governance actor, as their behaviors, such as reporting, blocking, or responding to harmful content, directly shape the visibility and diffusion of online messages (Helberger et al., 2018).

Second, governance interventions can be categorized as preventive (e.g., upload filters, before publication) or reactive (e.g., post-publication measures to mitigate impact and spread; Mündges, 2022). The present study focuses on reactive strategies, which counter harmful communication after its publication without suppressing it outright—an approach particularly relevant given concerns about freedom of expression.

Among platform-based reactive measures, content labels attach additional information to a post in order to contextualize it for users and support its evaluation in the form of fact-checks, sensitivity warnings, or notices providing additional contextual information (Morrow et al., 2022). Such labels constitute a comparatively light form of moderation because the original contribution remains accessible. Their effects may depend on their purpose, presentation, and specific implementation (Morrow et al., 2022). Research on misinformation suggests that warning labels can reduce perceived credibility and sharing intentions, although their effectiveness depends on their design and context (Mena, 2020). Whether similar effects occur for visual hate speech remains unclear.

User-centered interventions rely on communicative responses. Counter speech refers to responses that challenge, condemn, or reframe hateful content without restricting its visibility. It may address the original

communicator, express solidarity with the targeted group, introduce alternative interpretations, or signal to observing users that the hateful position is contested. Prior research shows that empathy-based textual counter speech can reduce support for hateful messages and discourage subsequent hate speech (Gennaro et al., 2025; Hangartner et al., 2021). Hangartner et al. (2021), for example, found consistent effects for empathy-oriented responses, but not for humor-based responses or warnings of consequences. Advances in GenAI add a further dimension. Wu et al. (2025) show that AI-generated textual counter speech can strengthen intentions to confront hate speech, and that empathy-oriented messages may be particularly promising. However, this evidence does not establish whether AI-generated visual responses produce similar effects.

Visual counter speech can challenge a hateful image by contesting its visual framing, appropriating its symbols, or presenting an alternative representation of the targeted group. Such responses may be particularly suitable for image-centered platform environments because they communicate through the same modality as the original message and, for formats such as memes, rely on recognizable templates, shared references, and the interplay of text and imagery (Carney & Levin, 2002; Schmid et al., 2024; Shifman, 2014). However, these characteristics do not necessarily mean that visual responses are more effective than textual ones.

2.1. Hypotheses and Research Questions on the Effectiveness of Governance Measures

A central assumption of the governance perspective is that interventions can shape user behavior and thereby influence how online content is endorsed, redistributed, opposed, or reported, and thus affect its diffusion dynamics (Helberger et al., 2018). The present study analyzes user engagement that is consequential for governance, thereby distinguishing amplification-oriented behaviors from oppositional responses, rather than relying on attitudinal proxies or platform-level diffusion data unavailable in an experimental setting. Liking expresses visible approval, sharing and external forwarding distribute content to additional audiences, and supportive commenting publicly reinforces its message. Together, these behaviors capture different forms of endorsement or (algorithmic) amplification. Critical commenting and reporting represent different forms of opposition. Critical comments publicly challenge the hateful message and may constitute counter speech. However, they do not necessarily reduce a post's visibility because engagement-based ranking systems may treat comments as signals of relevance regardless of their position toward the content (Milli et al., 2025). Reporting, by contrast, is a platform-directed request for review or moderation. Accordingly, reduced supportive commenting and increased critical commenting should not be treated as opposite ends of a single dimension.

Prior research indicates that both platform-based and user-centered interventions can reduce the spread of harmful or misleading information (Gennaro et al., 2025; Hangartner et al., 2021; Mena, 2020). In line with this literature, we therefore expect governance interventions to reduce amplification-oriented behaviors and increase oppositional responses:

H1: Individuals exposed to governance interventions will be less likely to (a) like, (b) share, (c) forward, or (d) supportively comment on hate images, and more likely to (e) critically comment on or (f) report such content compared to individuals in a control condition without intervention.

While governance measures are expected to have an overall mitigating effect, their effectiveness may vary depending on their modality. Research on visual communication suggests that images have particular advantages in attracting attention, communicating meaning, and remaining memorable (Carney & Levin, 2002; Leavitt, 2014). Moreover, the recognition and evaluation of hate speech vary according to whether it is presented textually or visually, and according to its level of directness (Schmid et al., 2024). A modality-congruence perspective suggests that a response may be particularly effective when it operates in the same communicative mode as the message it challenges. Research in brand communication indicates that congruence between the modality in which information is initially encountered and the modality of a subsequent message can facilitate processing and influence evaluations (Fransen et al., 2010). Applied to visual hate speech, counter-images may therefore be easier to relate directly to the original post than a separate textual response. This expectation may be especially relevant in image-centered platform environments, where, for example, hateful memes and counter-images share visual conventions such as recognizable templates, symbols, and combinations of text and imagery (Schmid et al., 2024). A visual response may consequently attract attention within the same communicative setting and make the relationship between the hateful message and its rejection more immediately apparent:

H2: Individuals exposed to counter-images will be less likely to (a) like, (b) share, (c) forward, or (d) supportively comment on hate images, and more likely to (e) critically comment on or (f) report such content compared to individuals exposed to textual counter speech or no intervention.

The present study further distinguishes generic from specific counter-images. Specific counter-images adopt or transform recognizable elements of the original hate image and thereby make the relationship between message and response explicit. Generic counter-images condemn hate more broadly without directly referring to the original content. They may be applicable across contexts but require users to infer their connection to the hateful post. As these forms have rarely been compared, their relative effectiveness remains an open question. GenAI can support the creation of both types of counter-images from comparatively simple inputs and thereby facilitate context-sensitive visual responses. However, the use of GenAI does not guarantee that a counter image will be persuasive, comprehensible, or appropriate. Specific counter-images may be more effective than generic counter-images because they directly reference and reframe recognizable elements of the original hate image, making the relationship between the harmful message and its rejection more immediately apparent. However, given the still limited body of empirical research on (AI-generated) visual counter speech, we formulate this expectation as a directional research question (RQ):

RQ1: Are specific counter-images that stylistically or textually respond to hate images more effective than generic counter-images in reducing amplification-oriented engagement (like, share, forward, supportive comment) and increasing oppositional engagement (critically comment, report) with hate images?

In addition, platform-issued labels and user-generated counter speech represent different governance sources. Labels communicate an institutional evaluation, whereas counter speech introduces an alternative voice within the online discussion and may be perceived as a more direct and socially meaningful rejection of the hateful message. However, as source, modality, and message content are partly intertwined in the interventions, the comparison is exploratory:

RQ2: Does the effectiveness of governance interventions in reducing amplification-oriented engagement (like, share, forward, supportive comment) and increasing oppositional engagement (critically comment, report) differ depending on whether they are attributed to platforms (e.g., labeling) or individual users (e.g., counter speech)?

Finally, the effectiveness of governance interventions may depend on characteristics of the hate content itself. Humorous hate speech combines degrading or exclusionary content with humor cues such as irony, parody, incongruity, or familiar meme conventions (Schmid, 2025). Humor does, however, not imply harmlessness or low intensity, but it can make hostility more ambiguous and reduce its immediate recognition (Schmid, 2025; Schmid et al., 2024), making visual hate speech less likely to be perceived as problematic and thus more likely to be shared. In contrast, overtly aggressive or violence-inciting content may trigger stronger negative reactions and be more readily recognized as inappropriate compared to humorous hate images (Schmid, 2025). As a result, baseline engagement patterns may differ depending on the intensity and framing of hate content:

H3: Individuals exposed to humorous hate images will be more likely to (a) like, (b) share, (c) forward, or (d) supportively comment on such content—and less likely to (e) critically comment on or (f) report it—compared to individuals exposed to aggressive or violence-inciting hate images, regardless of governance interventions.

3. Method

To examine the effects of different governance measures on user interactions with visual hate speech related to transgender identities and gender diversity, we conducted a standardized survey using an experimental design. Experimental approaches, relatively rare in media and communication policy research, allow for the systematic testing of governance options and the causal assessment of their effects (Puppis & Van den Bulck, 2024). The focus on transgender identities was chosen because transgender individuals are disproportionately targeted by online hate speech and discrimination, both in digital environments and beyond, making this issue a particularly salient and empirically relevant case for examining governance interventions (Oehmer-Pedrazzi & Pedrazzi, 2024a; Sánchez-Sánchez et al., 2024). We employed a 5×3 between-subjects factorial design. Participants were randomly assigned to one of 15 experimental conditions, resulting from the combination of five governance interventions and three levels of hate image intensity.

3.1. Experimental Conditions

Five governance conditions were tested:

1. Control condition (no governance intervention);
2. Generic counter image: A user-generated visual comment expressing general support for transgender individuals without directly referencing the hate image;
3. Specific counter image: A user-generated visual comment that explicitly engages with and counters elements of the hate image (both visually and textually);
4. Empathy-based textual counter speech: A user comment highlighting the harmful and hurtful nature of the post;
5. Platform-issued warning label: A content label indicating that the post may be offensive or harmful.

In addition, three levels of hate image intensity were operationalized:

1. Humorous/non-aggressive: Depictions that marginalize or exclude individuals or groups without using explicitly aggressive, insulting, dehumanizing, threatening, or violence-inciting visual or textual elements; this category also includes humor-based exclusion (e.g., memes);
2. Aggressive: Content that degrades, insults, or dehumanizes individuals or groups using explicit hostile imagery or language;
3. Violence-inciting: Content that endorses, legitimizes, or calls for violence against individuals or groups.

3.2. Procedure and Stimuli

Participants were recruited in the German-speaking part of Switzerland through the online panel provider Bilendi. Participation was voluntary, and respondents received a small financial incentive. Data collection took place between August 21 and August 30, 2024.

At the beginning of the survey, participants were exposed to two neutral, gender-unrelated social media posts in a simulated timeline and asked to indicate for each how likely they would be to engage in the six interaction behaviors (liking, sharing, forwarding, supportive/critical commenting, reporting); this familiarized participants with the response format and established a behavioral baseline.

Subsequently, participants were randomly assigned to one of the 15 experimental conditions and shown a third post containing one of three hate images targeting transgender individuals, varying in intensity (see Figure 1). The hate images were identified through a prior citizen science project conducted in Switzerland and were manually coded for intensity (Oehmer-Pedrazzi & Pedrazzi, 2024a).

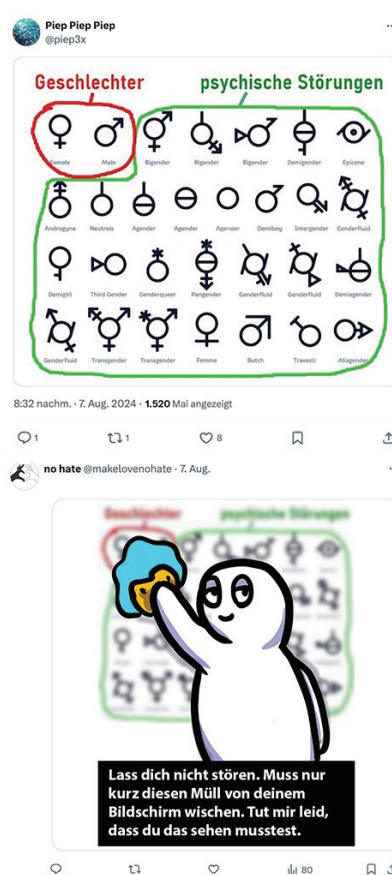
Depending on the experimental condition, the hate image was presented either without any intervention or combined with one of the four governance measures. In the case of counter speech and counter-images, the intervention appeared as a user comment posted by a fictitious account (“no hate @makelovenohate”). In the labelling condition, a researcher-designed warning notice with a light-orange background presented as platform-issued was displayed immediately above the post to indicate that the content could disparage or demean individuals or social groups (see Figure 1 for specific counter-images; Figure 2 for other interventions). All other elements, such as username, timestamp, and engagement metrics (likes, shares, views), were held constant across conditions to ensure internal validity.

The generic and specific counter-images were created by students from the joint Bachelor’s program Multimedia Production of the University of Applied Sciences of the Grisons and the Bern Academy of the Arts. The students were asked to produce, for each hate image, three specific counter-images (directly addressing it) and three generic ones (addressing its broader theme), each pursuing at least one of three goals: encouraging the sender’s reflection, discouraging further engagement with the hate image, or expressing solidarity with the targeted group. All counter-images were created using GenAI for idea development and/or image generation.

Humorous / non-aggressive post
with specific counter image



Aggressive post
with specific counter image



Violence inciting post
with specific counter image



Figure 1. Posts including hate images of varying intensity and specific counter-images used in the experiment. Notes: Left post, upper image—"I, too, was once a man trapped in a woman's body. Then I was born!"; left post, lower image—"Even Chuck Norris used to be a hater. Then he grew up"; middle post, upper image—"Genders" (left) and "Mental disorders" (right); middle post, lower image: "Don't let it bother you. Just need to wipe this garbage off your screen real quick. Sorry you had to see that."

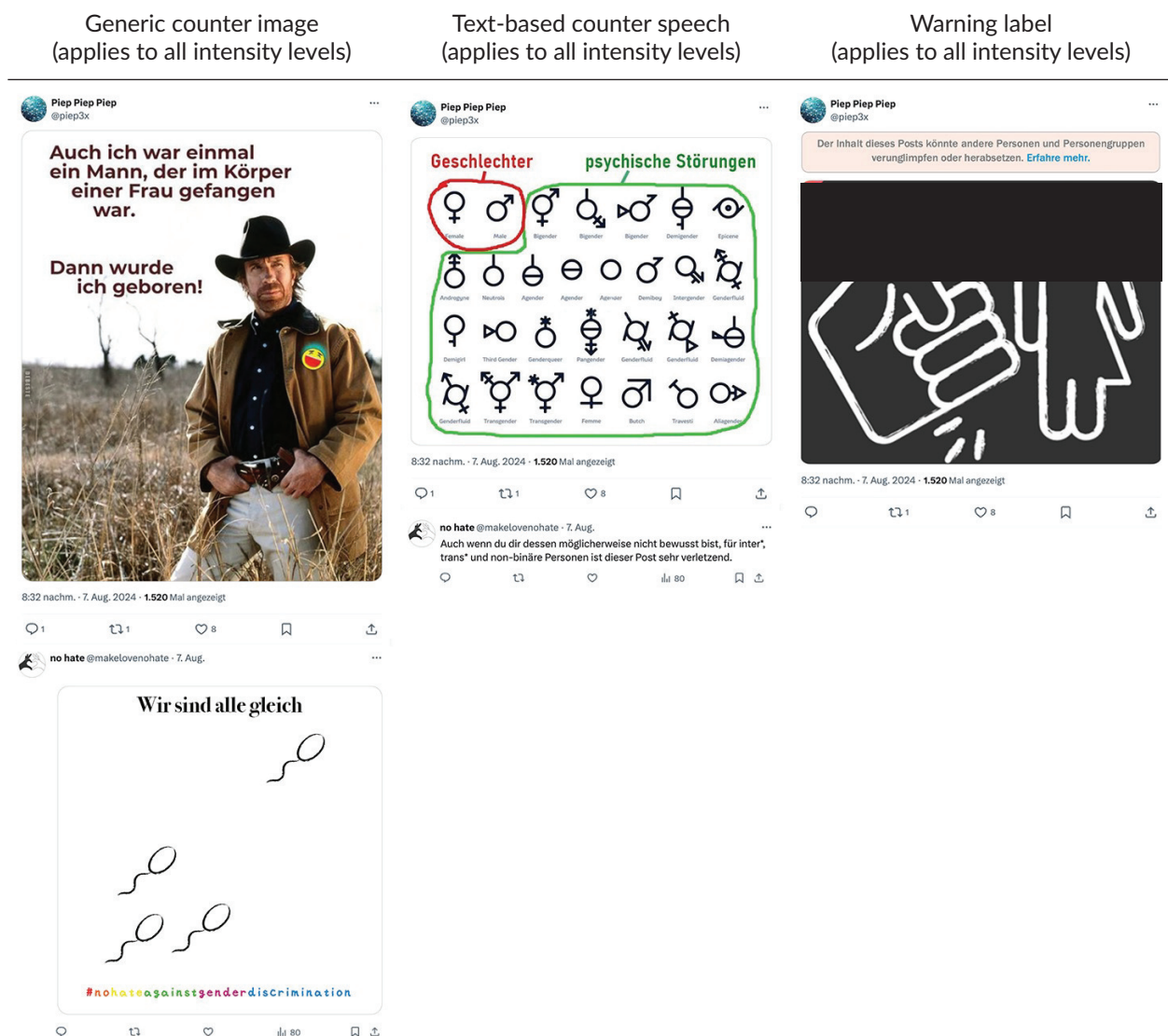


Figure 2. Posts including non-specific countermeasures used in the experiment. Notes: Left image—“We are all equal”; middle image—“Even if you may not be aware of it, this post is very hurtful to intersex, trans, and non-binary people”; right image—“The content of this post may disparage or demean other individuals or groups. Learn more.”

After exposure to the stimulus, participants again reported the likelihood of engaging with the post across several behavioral dimensions. No direct manipulation checks were included. Finally, relevant control variables were measured before the survey concluded with a debriefing and information about support services.

3.3. Ethical Considerations

Given that participants were exposed to potentially harmful content, several ethical precautions were implemented. Respondents gave explicit informed consent after being told about the study’s purpose, data handling, and the possibility of encountering offensive or distressing images. A welcome-page warning also provided contact information for support services, including national helplines and discrimination-counseling services in Bern (Switzerland). Furthermore, participants could withdraw at any time without penalty.

Participation was restricted to individuals aged 18 and older to protect minors from exposure to harmful content. No personally identifiable data were collected. Sociodemographic information (age, education, gender identity) was collected voluntarily. The study complied with applicable data protection regulations and was approved by the institutional review board of the Faculty of Economics and Social Sciences at the University of Fribourg (approval number: 2024-01-03). The study was preregistered on the Open Science Framework (Oehmer-Pedrazzi & Pedrazzi, 2024b).

3.4. Participants

An a priori power analysis using G*Power (Faul et al., 2009) indicated a required sample size of $N = 765$ to detect small to medium effects ($f = 0.175$) with a statistical power of 0.90 and an alpha level of 0.05 across 15 groups.

A total of 782 participants completed the survey. After excluding cases with implausibly short (<3 minutes) or long (>15 minutes) completion times, the final sample consisted of $N = 666$ respondents. The final sample therefore fell below the target for 90% power but still exceeded the sample size required for 80% power ($N = 615$). Group sizes ranged from $n = 40$ to $n = 48$.

Participants' ages ranged from 18 to 80 years ($M = 47.31$, $SD = 16.42$). Of the sample, 51.2% identified as female, 47.9% as male, 0.3% as diverse, and 0.6% chose not to disclose their gender. Approximately 39.9% had completed tertiary education. Randomization checks indicated no systematic differences across conditions in baseline engagement, age, gender identity, or education.

3.5. Measures

The dependent variables captured participants' intended engagement with the hate post. Participants were asked: "Auf einer Skala von 0–100 Prozent: Wie wahrscheinlich würden Sie den Ursprungspost von 'Piep Piep Piep'...?" (On a scale from 0% to 100%, how likely would you be to...the original post by the account "Piep Piep Piep"?). Using continuous sliders ranging from 0% to 100% (default position = 0%), participants indicated the likelihood that they would: like ($M = 18.54$, $SD = 31.07$); share publicly ($M = 12.23$, $SD = 24.60$); forward to friends or acquaintances ($M = 14.73$, $SD = 26.91$); comment supportively ($M = 9.70$, $SD = 22.23$); comment critically ($M = 12.75$, $SD = 26.21$); and report as inappropriate or impermissible ($M = 16.22$, $SD = 29.22$).

Continuous sliders were used because the items framed behavioral intentions as subjective probabilities ("how likely would you be to..."). Research suggests that survey-based sharing intentions correspond meaningfully with actual sharing behavior on social media (Mosleh et al., 2020).

To account for individual differences, several control variables were included, measured on five-point Likert scales:

- Attitudes toward online hate ($M = 4.42$, $SD = 0.85$, Spearman-Brown's $\rho = 0.81$): measured using two items assessing perceived severity and the need for stronger countermeasures;
- Attitudes toward transgender people and gender identity ($M = 3.54$, $SD = 1.10$, Cronbach's $\alpha = 0.90$): based on an adapted version of the scale by Walch et al. (2012);

- Social digital media literacy ($M = 4.20$, $SD = 0.82$, Cronbach's $\alpha = 0.84$): measured using items capturing respectful and responsible online communication behavior (Hoffmann et al., 2019);
- General engagement propensity: measured through participants' responses to the two neutral posts presented before the experimental stimulus. For each behavioral outcome, a baseline measure was computed by averaging the corresponding responses across the two neutral posts.

In addition, age, gender, and education were included as sociodemographic controls.

The exact wording of all items, including the control-variable scales and the six dependent-variable prompts, is reported in the Supplementary Material.

3.6. Data Analysis

The analysis strategy deviates slightly from the preregistered plan. While the preregistration proposed analyses of variance (ANOVA), we instead estimated baseline-adjusted linear regression models with planned contrasts based on estimated marginal means. Both approaches are based on the general linear model, but the regression framework provided greater flexibility for incorporating covariates and testing theoretically specified contrasts. The deviation reflects a refinement of the analytical strategy and does not affect the hypotheses, operationalizations, or interpretation of the results.

To examine differences between governance interventions, we conducted planned contrasts comparing theoretically relevant groups of interventions, including visual versus text-based approaches as well as specific versus generic counter-images. Given directional expectations, one-tailed tests were applied.

All models were estimated using ordinary least squares regression with HC3 heteroskedasticity-robust standard errors. Control variables included baseline engagement, age, gender, education, social media literacy, and attitudinal indices.

4. Results

4.1. Overall Effects of Governance Interventions

Across behavioral outcomes, governance interventions showed a selective pattern of effects. Compared with the control condition, the interventions reduced the stated likelihood of forwarding the post by an average of 5.59 points on the 0–100 scale ($b = -5.59$, $p = 0.013$, one-tailed). The estimated effects on sharing, liking, and supportive commenting also pointed in the hypothesized negative direction, but did not reach statistical significance. By contrast, the interventions did not significantly increase oppositional behaviors, either in terms of critical commenting or reporting (see Figure 3). Overall, these findings provide partial support for H1. The effects were more evident for reducing the dissemination of the content than for encouraging users to respond actively against it.

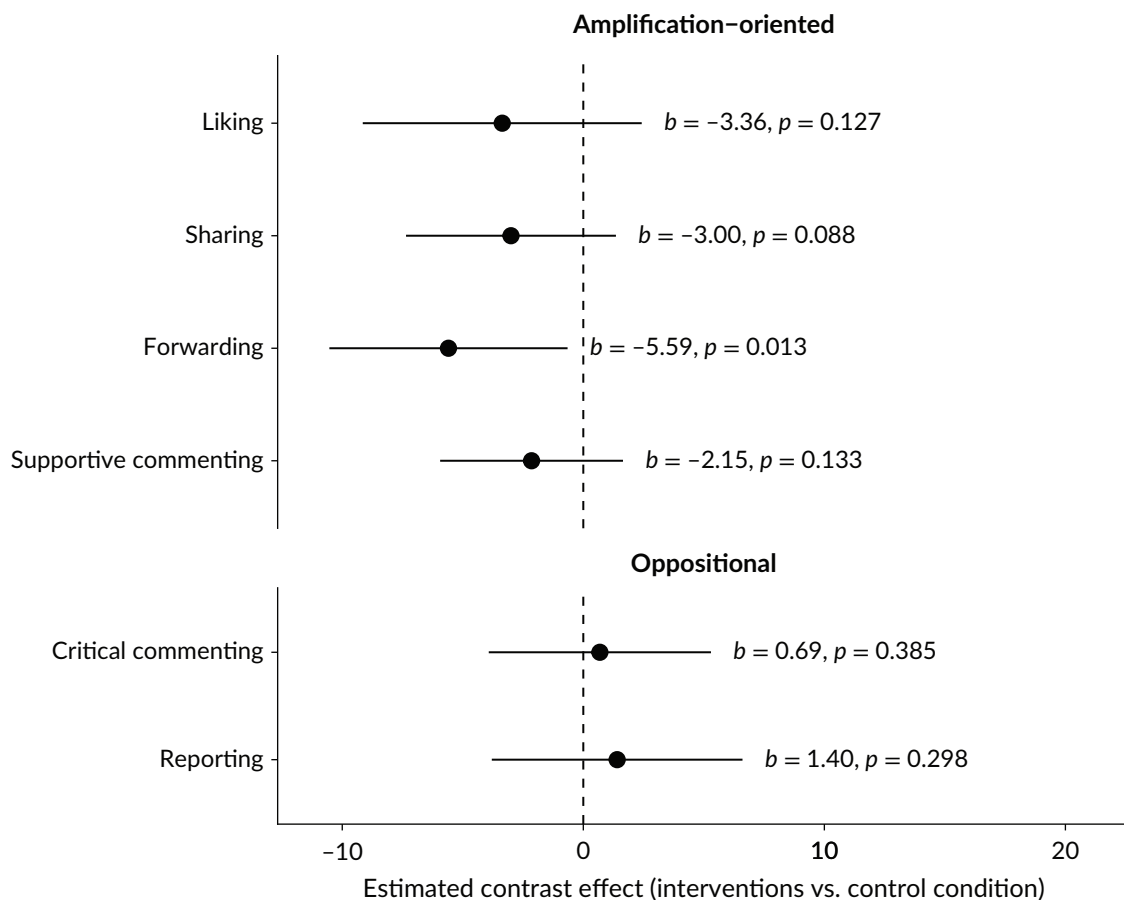


Figure 3. Overall effects of governance interventions across user responses. Notes: Baseline-adjusted planned contrasts comparing all governance intervention conditions to the control condition across behavioral outcomes; points represent estimated contrasts, and error bars indicate 95% confidence intervals based on HC3 heteroskedasticity-consistent standard errors; the reported *p*-values are one-tailed; positive values indicate higher levels of the respective behavior compared to the control condition, whereas negative values indicate lower levels; *N* = 630 (due to listwise deletion of missing values).

4.2. Differences Between Governance Interventions

Next, we compared visual interventions (generic and specific counter-images) with empathic counter speech and with the control condition. Across outcomes, visual interventions did not consistently outperform textual counter speech (all $p > 0.68$). Relative to the control condition, visual interventions significantly reduced forwarding ($b = -5.29$, $p = 0.029$, one-tailed) and pointed in the expected direction for sharing ($b = -3.19$, $p = 0.092$, one-tailed), with no significant effects on liking, supportive commenting, or oppositional behaviors (see Figure 4). Overall, these findings do not support H2, indicating that visual interventions do not outperform textual counter speech.

To address RQ1, we compared specific and generic counter-images. Across amplification-oriented behaviors, specific counter-images were more effective than generic ones. Significant reductions were observed for liking ($b = -6.69$, $p = 0.030$, one-tailed), forwarding ($b = -6.92$, $p = 0.014$, one-tailed), and supportive commenting ($b = -4.78$, $p = 0.024$, one-tailed). The effect on sharing pointed in the expected direction but did not reach statistical significance. No differences emerged for oppositional outcomes. These findings suggest that the

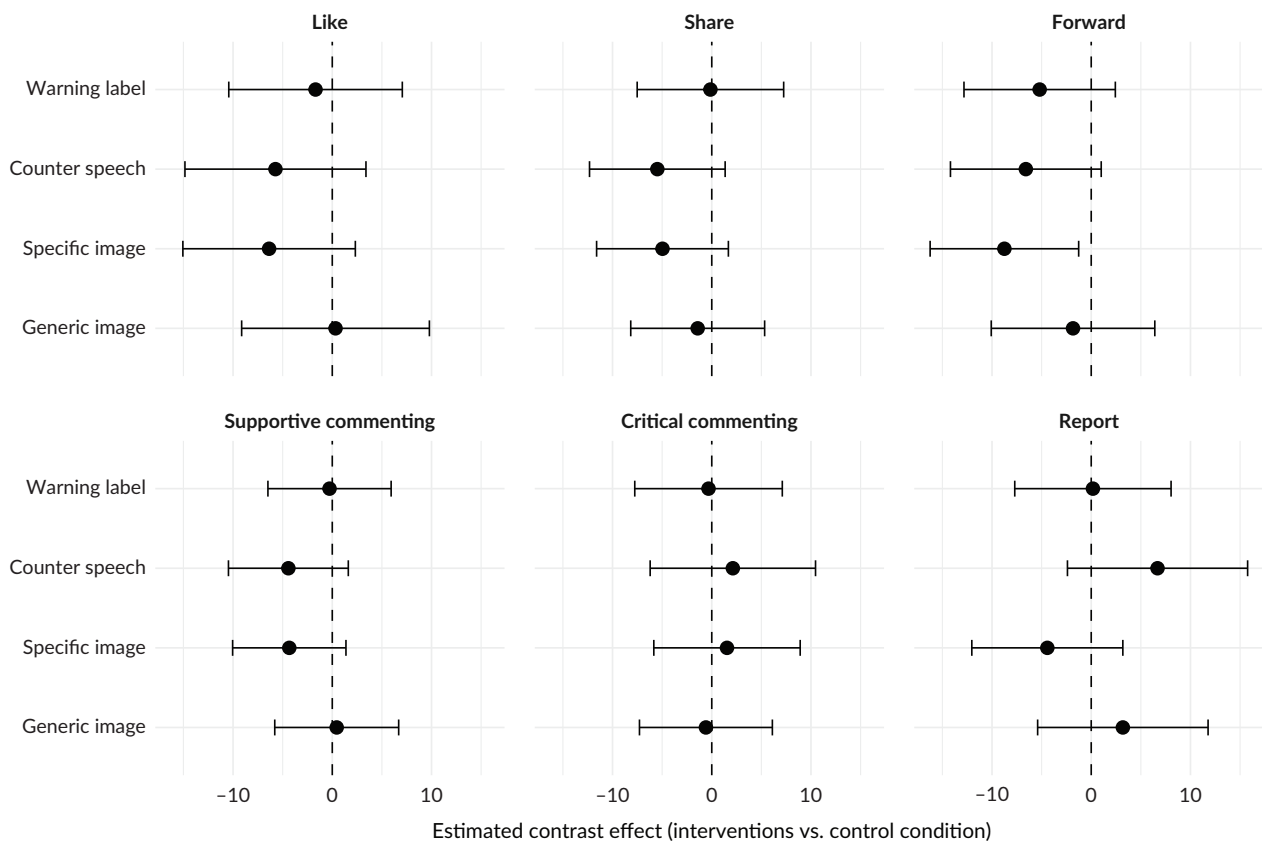


Figure 4. Effects of governance interventions on user responses. Notes: Points represent baseline-adjusted estimated contrasts comparing each intervention condition with the control condition; horizontal lines indicate 95% confidence intervals; positive estimates indicate higher levels of the respective behavior compared to the control condition, whereas negative estimates indicate lower levels; $N = 630$ (due to listwise deletion of missing values).

effectiveness of counter-images depends on their specificity and alignment with the original content, rather than on their visual format per se.

To address RQ2, we compared user-based interventions (generic and specific counter-images, empathic counter speech) with platform-based interventions (warning labels). Differences were largely absent: only sharing showed a marginal tendency favoring user-based interventions ($b = -3.82$, $p = 0.066$), with no significant differences for the remaining outcomes. Overall, these findings suggest that the source of the intervention—whether originating from users or the platform—plays only a limited role in shaping user responses.

Finally, we examined differences between individual intervention types by comparing each condition directly to the control group. Across outcomes, only specific counter-images showed a statistically significant effect, reducing forwarding intentions ($b = -8.75$, $p = 0.008$, one-tailed). For empathic counter-speech, the observed effects were in the expected direction for sharing, forwarding, and reporting, but not statistically significant (all $p < 0.14$). Generic counter-images and warning labels did not produce significant differences compared to the control condition (see Figure 5).

Overall, the results indicate that governance interventions primarily influence more effortful forms of content dissemination. Importantly, effectiveness appears to depend less on whether an intervention is visual or textual, and more on its specificity and contextual alignment with the content.

4.3. Effects of Hate Image Intensity

We further examined whether the effectiveness of governance interventions varies depending on the intensity of the hate image by conducting planned contrasts comparing humorous content with aggressive and violence-inciting content, averaged across governance interventions.

Across amplification-oriented behaviors, humorous content consistently elicited higher levels of engagement compared to violence-inciting content. Significant differences were observed for liking ($b = 10.73$, $p < 0.001$, one-tailed), sharing ($b = 5.78$, $p = 0.004$), and forwarding ($b = 9.77$, $p < 0.001$), while no significant differences emerged for supportive commenting. Comparisons between humorous and aggressive content showed no statistically significant differences for amplification-oriented behaviors, although estimates pointed in the expected direction. For oppositional responses, exposure to humorous content led to significantly lower reporting intentions compared to both aggressive ($b = -9.80$, $p < 0.001$) and violence-inciting content ($b = -9.71$, $p < 0.001$), as well as significantly lower critical commenting (aggressive: $b = -3.91$, $p = 0.038$; violence-inciting: $b = -6.51$, $p = 0.004$).

Overall, these findings provide strong support for H3. Humorous hate content elicits higher levels of engagement and lower levels of opposition compared to more severe forms of hate content, particularly when contrasted with violence-inciting material.

Additional exploratory analyses examined whether the effectiveness of individual governance interventions varied across levels of content intensity by comparing each intervention condition to the control group within each intensity level. Results revealed a differentiated pattern of effects across intensity levels (see Figure 5). For humorous content, specific counter-images significantly reduced sharing ($b = -12.70$, $p = 0.040$) and forwarding ($b = -17.50$, $p = 0.012$), while warning labels also reduced forwarding ($b = -14.39$, $p = 0.041$). No significant effects emerged for aggressive content. For violence-inciting content, empathic counter speech significantly reduced sharing ($b \approx -9.96$, $p = 0.009$) and supportive commenting ($b \approx -7.49$, $p = 0.034$). Across oppositional behaviors, no intervention led to statistically significant increases in critical commenting or reporting within any intensity level.

Overall, these exploratory findings suggest that the effectiveness of governance interventions depends on the intensity of hate content. Interventions appear most effective in reducing engagement with humorously framed hate, while for highly severe content, effects are more limited and primarily observable for empathic counter speech.

4.4. Robustness Check

As a robustness check, we additionally estimated a system of seemingly unrelated regression equations across all six outcomes. Residual correlations were substantial among spread-oriented behaviors (e.g., liking, sharing, and forwarding: $r \approx 0.66$ – 0.69) and between the two oppositional responses (critical commenting

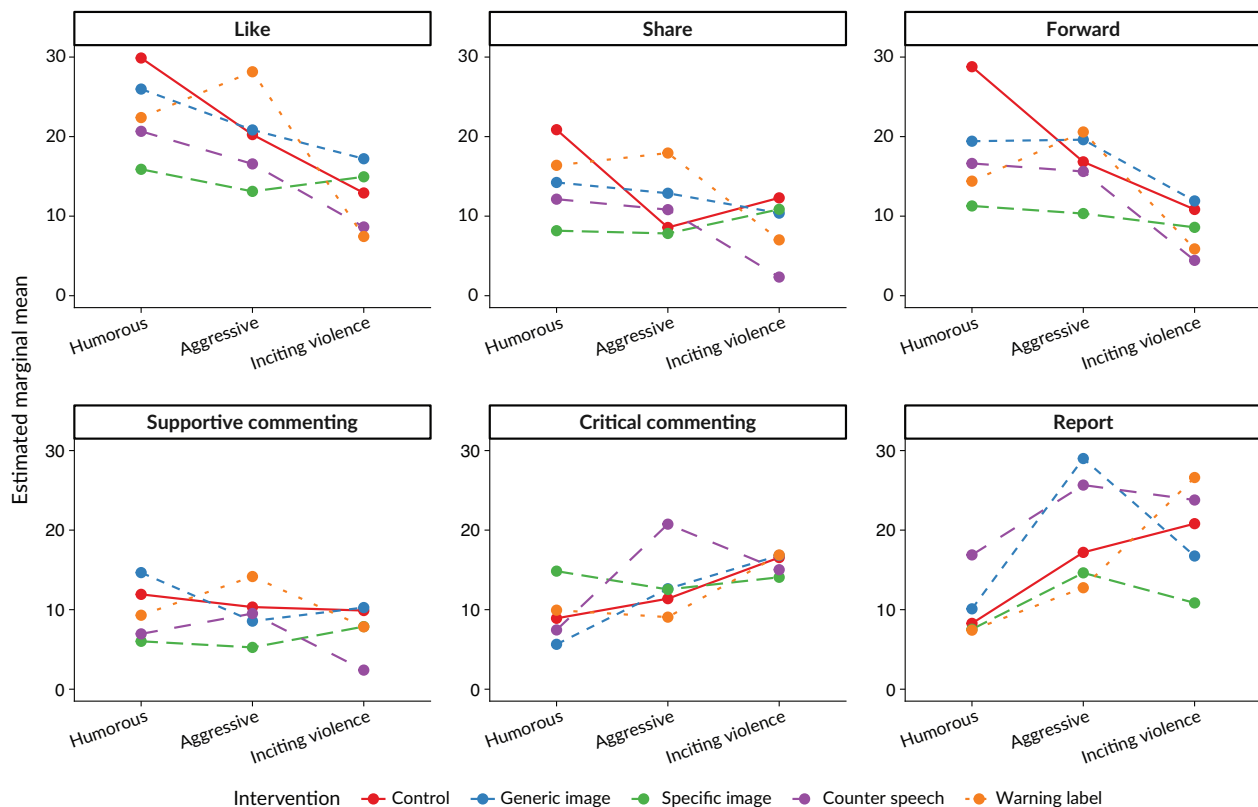


Figure 5. Predicted user responses by governance intervention and hate content intensity. Notes: Estimated marginal means for user engagement across behavioral outcomes (liking, sharing, forwarding, supportive commenting, critical commenting, and reporting) by governance intervention and content intensity (humorous, aggressive, violence-inciting); values are based on baseline-adjusted regression models with HC3 heteroskedasticity-robust standard errors; lower values indicate reduced engagement for amplification-oriented behaviors, whereas higher values indicate increased engagement for oppositional behaviors; $N = 630$ (due to listwise deletion of missing values).

and reporting: $r \approx 0.56$), indicating that these responses are empirically interrelated. The estimates from the seemingly unrelated regression system were substantively consistent with the main analyses, indicating that specific counter-images showed the clearest reductions in spread-related engagement, particularly for liking, sharing, and forwarding, whereas oppositional outcomes remained comparatively unaffected.

5. Conclusion

The present study examined the effectiveness of reactive governance interventions in shaping users' intended responses to visual hate speech, focusing on measures that operate ex post and therefore do not directly interfere with freedom of expression. By experimentally testing textual counter speech, generic and specific counter-images, and a platform-issued warning label, this study provides empirical insights into how different governance approaches shape user interactions with hate images targeting transgender individuals.

Overall, the findings demonstrate that governance interventions can indeed influence user behavior. Participants exposed to any form of intervention were less likely to share and forward hate images, with statistically robust effects for forwarding and consistent directional patterns for sharing compared to those in the control condition. This reduction in diffusion is particularly relevant given that privately forwarded

content often originates from trusted interpersonal sources and may therefore exert a stronger persuasive impact (Metzger et al., 2010). From a governance perspective, limiting such forms of redistribution represents a meaningful mechanism for curbing the broader circulation of harmful visual content.

Among the tested interventions, empathy-based textual counter speech and specific counter-images showed the most consistent patterns in reducing amplification-oriented behaviors, including liking, sharing, forwarding, and supportive commenting. Notably, and contrary to expectations derived from visual communication research, visual counter-images were not more effective than textual counter speech. The comparatively consistent performance of empathy-based textual counter speech also aligns with previous research showing that empathy-oriented messages can reduce subsequent hate speech or willingness to support it (Gennaro et al., 2025; Hangartner et al., 2021). One tentative explanation is that text can state the rejection of a hateful message explicitly and may therefore require less interpretive work than a visual response. However, the study did not directly measure participants' comprehension or interpretation of the interventions.

The comparison between generic and specific counter-images further showed that specificity, not visual format per se, drove effectiveness: Specific images produced stronger reductions across several amplification-oriented outcomes. This suggests that visual counter speech should not be understood merely as counter speech presented in image form—its meaning emerges relationally, through how it adopts, transforms, or rejects the symbols and conventions of the original hate image. Just as the interpretation of visual hate speech depends on presentation form, directness, and context (Schmid et al., 2024), the same appears to hold for visual responses: Specific counter-images make the relationship to the original content visible, whereas generic images require users to infer it.

Platform-issued warning labels, in contrast, showed weaker and less consistent effects: Although they reduced forwarding in some contexts, they were generally less effective than user-generated interventions, and for aggressive content even showed occasional counterproductive tendencies—possibly reflecting user skepticism toward platform moderation (Ozanne et al., 2022). This is consistent with the broader labeling literature, which shows that label effectiveness is highly sensitive to design and wording and that most existing evidence stems from misinformation rather than hate speech contexts (Morrow et al., 2022); the single, generic warning-label design tested here may therefore understate the potential of platform labeling more broadly.

The effectiveness of governance interventions further depended on the intensity of the hate content. Humor-based hate images were more likely to be endorsed and shared, whereas aggressive and especially violence-inciting images elicited less supportive engagement and were more readily rejected—consistent with research showing that humor cues and familiar meme conventions make hateful meanings more ambiguous and harder to recognize as hostile (Schmid, 2025; Schmid et al., 2024). Governance interventions were correspondingly most effective for humorous, more ambiguously harmful content, and less effective as the content's harmful nature became more obvious.

Across all analyses, governance interventions did not consistently increase oppositional responses such as critical commenting or reporting, indicating that the measures primarily discouraged engagement rather than actively mobilizing counteraction.

These findings carry several implications for the governance of visual hate speech. First, they underscore the crucial role of users as active agents in shaping online discourse. User-generated responses—particularly empathetic counter speech and context-specific visual counter messages—can meaningfully reduce the spread of harmful content. By lowering engagement rates, such interventions may also indirectly influence algorithmic amplification processes, which typically prioritize highly interactive content. In this way, user-centered governance can contribute to fostering a less discriminatory and more constructive online environment. Advances in generative AI expand the potential of such user-driven interventions by enabling the rapid and low-cost creation of counter speech and counter-images that are adapted to the content of a particular hate post, thereby lowering barriers to participation and facilitating scalable forms of engagement in countering harmful visual content. As a practical illustration, a prompt for generating a specific counter image might read:

You have come across the attached hate image in your social media timeline. You want to clearly distance yourself from its message and express solidarity with those affected. Design a specific counter image that deliberately picks up central visual, linguistic, and stylistic features of the original and critically reinterprets them, so the connection to this specific hate image is immediately recognizable. The result should be professionally designed and suitable for publication on social media.

Second, enabling such user-driven interventions requires appropriate platform design and digital literacy. Platforms should expand affordances that let users respond not only with text but also with visual content, as this would enable more direct engagement with visual hate speech. At the same time, educational initiatives are needed to raise awareness of the impact of user interactions and to empower individuals to actively counter harmful content.

Third, independent “trusted commentators” who systematically engage in counter speech and GenAI-assisted counter-image creation may represent a promising governance approach, provided appropriate accountability and transparency mechanisms are in place.

Finally, the findings suggest a differentiated use of platform-based interventions. Warning labels should be applied cautiously and may be most appropriate for highly severe forms of visual hate, while overuse in less severe cases may risk unintended effects. At the same time, the challenges of automated content moderation—particularly in interpreting complex visual and multimodal content—remain substantial despite advances in AI-based systems (Cheema et al., 2023).

Several limitations should be acknowledged. First, the study relies on self-reported behavioral intentions rather than observed behavior, which may not fully reflect actual user actions in real-world settings. However, prior research suggests that such measures provide a reasonable approximation of behavior (Mosleh et al., 2020). Relatedly, the absence of a manipulation check means we cannot directly verify that participants consciously registered the intervention or intensity level to which they were exposed. Second, this study tested only a single design of warning label. Given evidence that labeling effects depend strongly on wording, framing, and visual presentation (Morrow et al., 2022), the comparatively weak performance of warning labels observed here should not be generalized to platform labeling as a whole, and future research should systematically vary label design. Third, the findings are based on a limited set of hate images targeting transgender individuals. Future research should test generalizability to other target groups, visual

content, and intensity levels. Qualitative approaches, such as interviews or think-aloud protocols (Bilandzic, 2017), would further illuminate how users interpret and respond to governance interventions.

In sum, this study highlights the potential of AI-generated or AI-assisted counter-images as a reactive governance intervention in addressing visual hate speech. The findings suggest that the effectiveness of counter speech cannot be inferred from its modality alone. Instead, contextual specificity, the clarity of the response, and the framing of the hate content appear to shape whether an intervention affects users' intended behavior. AI-generated specific counter-images are therefore promising not because they are visual or AI-produced, but because GenAI can support the creation of responses that directly engage with the visual meaning of a particular hate message. Together with the stronger effects observed for ambiguously harmful, humor-based content, this contributes to a more differentiated understanding of when and how visual counter speech may complement platform and user-centered efforts to govern online hate.

Funding

The study was funded by the Swiss Federal Office of Communications OFCOM.

Conflict of Interests

The authors declare no conflicts of interest.

LLMs Disclosure

An LLM (ChatGPT, OpenAI) was used to assist with language editing and translation. All outputs were carefully reviewed, revised, and approved by the authors, who retain full responsibility for the content of this manuscript.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

- Arcila Calderón, C., Sánchez Holgado, P., Gómez, J., Barbosa, M., Qi, H., Matilla, A., Amado, P., Guzmán, A., López-Matías, D., & Fernández-Villazala, T. (2024). From online hate speech to offline hate crime: The role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanities and Social Sciences Communications*, 11(1), Article 1369. <https://doi.org/10.1057/s41599-024-03899-1>
- Bilandzic, H. (2017). Lautes denken. In L. Mikos & C. Wegener (Eds.), *Qualitative Medienforschung: Ein Handbuch* (pp. 406–413). UVK.
- Bin Aziz, J., & Briel, H. (2024). Visual hate speech and its discontents: Young adult Chinese and Malaysians' perception of visual hate speech. *International Journal of Communication*, 18, 27–47. <https://ijoc.org/index.php/ijoc/article/view/19787>
- Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, 6(4), 58–69. <https://doi.org/10.17645/mac.v6i4.1493>
- Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14, 5–26. <https://doi.org/10.1023/A:1013176309260>
- Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech: Systematic review. *Aggression and Violent Behavior*, 58, Article 101608. <https://doi.org/10.1016/j.avb.2021.101608>

- Cheema, G. S., Hakimov, S., Müller-Budack, E., Otto, C., Bateman, J. A., & Ewerth, R. (2023). Understanding image-text relations and news values for multimodal news analysis. *Frontiers in Artificial Intelligence*, 6, Article 1125533. <https://doi.org/10.3389/frai.2023.1125533>
- Crawford, B., Keen, F., & Suarez-Tangil, G. (2021). Memes, radicalisation, and the promotion of violence on chan sites. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 982–991. <https://doi.org/10.1609/icwsm.v15i1.18121>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fransen, M. L., Fennis, B. M., & Pruyn, A. T. H. (2010). Matching communication modalities: The effects of modality congruence and processing style on brand evaluation and brand choice. *Communication Research*, 37(4), 576–598. <https://doi.org/10.1177/0093650210368251>
- Gennaro, G., Derksen, L., Abdelrahman, A., Brogginì, E., Green, M. A., Haerter, V. A., Heer, E., Heidler, I., Kauer, F., Kim, H.-N., Landry, B., Levis, A., Li, J., Şimşir, Ş., Srbínovska, I., Vital, R. A., Donnay, K., Gilardi, F., & Hangartner, D. (2025). Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media. *Scientific Reports*, 15(1), Article 22018. <https://doi.org/10.1038/s41598-025-05041-w>
- Hangartner, D., Gennaro, G., Alasiri, S., Bahrigh, N., Bornhoft, A., Boucher, J., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50), Article e2116310118. <https://doi.org/10.1073/pnas.2116310118>
- Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, 34(1), 1–14. <https://doi.org/10.1080/01972243.2017.1391913>
- Hietanen, M., & Eddebo, J. (2023). Towards a definition of hate speech—With a focus on online contexts. *Journal of Communication Inquiry*, 47(4), 440–458. <https://doi.org/10.1177/01968599221124309>
- Hoffmann, C. P., Weber, J., Zepic, R., Greger, V., & Krcmar, H. (2019). Dimensionen digitaler Mündigkeit und politische Beteiligung im Netz. In I. Engelmann, M. Legrand, & H. Marzinkowski (Eds.), *Politische Partizipation im Medienwandel* (pp. 79–99). Nomos. <https://doi.org/10.17174/dcr.v6.4>
- Hornuff, D. (2020). *Hassbilder: Gewalt posten, Erniedrigung liken, Feindschaft teilen*. Wagenbach.
- Knobloch, S., Hastall, M., Zillmann, D., & Callison, C. (2003). Imagery effects on the selective reading of internet newsmagazines. *Communication Research*, 30(1), 3–29. <https://doi.org/10.1177/0093650202239023>
- Leavitt, A. (2014). From #FollowFriday to YOLO: Exploring the cultural salience of Twitter memes. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 137–154). Peter Lang.
- Marquart, F. (2023). Video killed the Instagram star: The future of political communication is audio-visual. *Journal of Visual Political Communication*, 10(1), 49–57. https://doi.org/10.1386/jvpc_00024_1
- Mayntz, R. (2004). Governance im modernen Staat. In A. Benz (Ed.), *Governance—Regieren in komplexen Regelsystemen: Eine Einführung* (pp. 65–76). Springer. https://doi.org/10.1007/978-3-531-90171-8_4
- Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, 12(2), 165–183. <https://doi.org/10.1002/poi3.214>
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), Article pgaf062. <https://doi.org/10.1093/pnasnexus/pgaf062>

- Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10), 1365–1386. <https://doi.org/10.1002/asi.24637>
- Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS ONE*, 15(2), Article e0228882. <https://doi.org/10.1371/journal.pone.0228882>
- Mündges, S. (2022). Die Verantwortung von Facebook & Co.: Über den Umgang digitaler Plattformen mit Hate Speech. In G. Weitzel & S. Mündges (Eds.), *Hate speech: Aktivismus—und Propagandaforschung* (pp. 231–248). Springer. https://doi.org/10.1007/978-3-658-35658-3_12
- Oehmer-Pedrazzi, F., & Pedrazzi, S. (2024a). “An image hurts more than 1000 words?” Sources, channels, and characteristics of digital hate images. *Communications*, 49(3), 421–443. <https://doi.org/10.1515/commun-2023-0117>
- Oehmer-Pedrazzi, F., & Pedrazzi, S. (2024b). *Governance of visual hate images on the internet*. OSF. <https://osf.io/urbvs>
- Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, 9(2). <https://doi.org/10.1177/20539517221115666>
- Puppis, M. (2010). Media governance: A new concept for the analysis of media policy and regulation. *Communication, Culture & Critique*, 3(2), 134–149. <https://doi.org/10.1111/j.1753-9137.2010.01063.x>
- Puppis, M., & Van den Bulck, H. (2024). Methods for global media and communication governance research. In C. Padovani, V. Wavre, A. Hintz, G. Goggin, & P. Iosifidis (Eds.), *Global communication governance at the crossroads* (pp. 371–387). Springer. https://doi.org/10.1007/978-3-031-29616-1_21
- Sánchez-Sánchez, F., Ruiz-Muñoz, D., & Sánchez-Sánchez, F. J. (2024). Mapping homophobia and transphobia on social media. *Sexuality Research and Social Policy*, 21, 210–226. <https://doi.org/10.1007/s13178-023-00879-z>
- 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., 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>
- Schünemann, W. J., & Steiger, S. (2023). Die Regulierung von Internetinhalten am Beispiel Hassrede: Ein Forschungsüberblick. In S. Jaki & S. Steiger (Eds.), *Digitale hate speech* (pp. 155–171). J. B. Metzler. https://doi.org/10.1007/978-3-662-65964-9_8
- Shabad, R., & Haake, G. (2026, February 6). Trump says he won’t apologize for racist post depicting the Obamas as apes. *NBC News*. <https://www.nbcnews.com/politics/white-house/trump-shares-racist-video-depicting-obamas-monkeys-rcna257756>
- Shifman, L. (2014). The cultural logic of photo-based meme genres. *Journal of Visual Culture*, 13(3), 340–358. <https://doi.org/10.1177/1470412914546577>
- Stahel, L., & Baier, D. (2023). Digital hate speech experiences across age groups and their impact on well-being: A nationally representative survey in Switzerland. *Cyberpsychology, Behavior, and Social Networking*, 26(7), 519–526. <https://doi.org/10.1089/cyber.2022.0185>
- Unger, D., & Unger-Sirsch, J. (2023). „Ihr gehört nicht dazu!“ Soziale Ausgrenzung durch Hate Speech als Problem für liberale Demokratien. In S. Jaki & S. Steiger (Eds.), *Digitale hate speech* (pp. 173–195). J. B. Metzler. https://doi.org/10.1007/978-3-662-65964-9_9

- Walch, S. E., Ngamake, S. T., Francisco, J., Stitt, R. L., & Shingler, K. A. (2012). The attitudes toward transgendered individuals scale: Psychometric properties. *Archives of Sexual Behavior*, 41(5), 1283–1291. <https://doi.org/10.1007/s10508-012-9995-6>
- Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social media? *Telematics and Informatics*, 101, Article 102304. <https://doi.org/10.1016/j.tele.2025.102304>
- Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. *Journal of King Saud University Computer and Information Sciences*, 37(6), Article 148. <https://doi.org/10.1007/s44443-025-00154-8>

About the Authors



Stefano Pedrazzi is a senior researcher and lecturer in communication science at the University of Fribourg, Switzerland.



Franziska Oehmer-Pedrazzi is a professor of communication science at the University of Applied Sciences of the Grisons, co-founder of the MILEVA Institut, and co-vice president of the Swiss Association of Communication and Media Research (SACM).