



Citation: Sgobba, E., & Scardigno, A. F. (2026). Spectatorship, bodies and digital hate: observing online hate speech and its implications for media education. *Media Education* 17(1): 67-81. doi: 10.36253/me-19844

Received: February, 2026

Accepted: April, 2026

Published: May, 2026

© 2026 Author(s). This is an open access, peer-reviewed article published by Firenze University Press (<https://www.fupress.com>) and distributed, except where otherwise noted, under the terms of the CC BY 4.0 License for content and CC0 1.0 Universal for metadata.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Spectatorship, bodies and digital hate: observing online hate speech and its implications for media education

Spettatorialità, corpi e odio digitale: un'analisi dell'hate speech online e delle sue implicazioni per la media education

ENRICA SGOBBA*, ANNA FAUSTA SCARDIGNO

Università di Bari, Italy

sgobba.enrica@gmail.com; annafausta.scardigno@uniba.it

*Corresponding author

Abstract. In hyperconnected online environments, hate speech has become a structural component of digital communication, deeply affecting processes of identity construction, social interaction and citizenship. This article investigates online hate speech by shifting the focus from perpetrators and victims to digital spectatorship, understood as a critical yet often passive position within platform-mediated interactions. Drawing on media education and realistic evaluation frameworks, the study explores how users observe, interpret and respond to discriminatory discourse targeting bodies and identities that deviate from dominant norms. Empirical data were collected through a mixed-method evaluative survey administered to Italian social media users, combining quantitative measures with qualitative open-ended responses, and through the Nominal Group Technique involving a panel of experts with academic, professional and activist experience in the field of online hate speech. Findings reveal a high exposure to hate speech, widespread non-intervention and the presence of moral disengagement mechanisms. The article argues that countering online hate speech requires moving beyond protective approaches toward a media education framework oriented to responsibility, emotional awareness and active digital citizenship. The study outlines educational implications aimed at fostering critical autonomy, communicative responsibility and inclusive participation in digital public spaces.

Keywords: hate speech, digital spectatorship, realistic evaluation, media education, moral disengagement.

Riassunto. Negli ambienti online iperconnessi, l'hate speech è diventato una componente strutturale della comunicazione digitale, incidendo profondamente sui processi di costruzione dell'identità, di interazione sociale e di cittadinanza. Questo articolo indaga l'hate speech online spostando l'attenzione dagli autori e dalle vittime alla spettatorialità digitale, intesa come una posizione critica ma spesso passiva all'interno delle interazioni mediate dalle piattaforme. Facendo riferimento ai quadri teorici della media education e della realistic evaluation, lo studio esplora il modo in cui gli utenti osservano, interpretano e rispondono ai discorsi discriminatori rivolti a corpi e identità.

tà che deviano dalle norme dominanti. I dati empirici sono stati raccolti attraverso una survey valutativa con metodo misto somministrata a utenti italiani dei social media, e attraverso la Nominal Group Technique, che ha coinvolto un panel di esperti con esperienza accademica, professionale e attivista nel campo dell'hate speech online. I risultati evidenziano un'elevata esposizione all'hate speech, una diffusa tendenza al non-intervento e la presenza di meccanismi di disimpegno morale. L'articolo sostiene che il contrasto all'hate speech online richiede di superare approcci meramente protettivi, orientandosi verso un modello di media education fondato sulla responsabilità, sulla consapevolezza emotiva e sulla cittadinanza digitale attiva. Lo studio delinea inoltre implicazioni educative volte a promuovere autonomia critica, responsabilità comunicativa e partecipazione inclusiva negli spazi pubblici digitali.

Parole chiave: hate speech, spettatorialità digitale, realistic evaluation, media education, disimpegno morale.

1. INTRODUCTION

In today's hyperconnected world, in which individuals act and construct meaning, the traditional distinction between online and offline is increasingly inadequate for describing the complexity of social and communicative practices. As noted by Floridi (2015), the neologism «onlife» captures the dynamic interconnection between the digital and the material dimensions, conveying the idea of a unified living environment in which information and communication technologies profoundly affect the human condition. These technologies do not merely mediate communication; they also reorganise spatial and temporal experience, as well as ways of relating to oneself, to others and to the world. In doing so, they challenge established paradigms and produce hybrid forms of subjectivity and identity.

Within this scenario, the Self, understood as a dynamic and relational social construction, is constantly negotiated through mediated interactions that are amplified and regulated by digital platforms and non-neutral socio-technical devices (Masterman, 1985). In these contexts, the boundaries between public and private are redefined, and the 'dramaturgical' dynamics of self-presentation intensify (Goffman, 1956). It is within this communicative arena that the growing manifestations of online hate speech emerge, a phenomenon that has acquired crucial educational relevance, especially with the advent of social media. Discursive hate, motivated by forms of discrimination related to ethnic origin, gender, sexual orientation, religion or disability, cannot be understood or addressed without considering the structuring role of media as environments of life and as «philosophical devices» that shape cultural universes and individuals' relationships with reality, with themselves and with others (Cappello, 2010).

From this perspective, media education¹ emerges as a fundamental theoretical and practical field for promot-

ing critical, reflective and participatory competences that support the exercise of responsible and informed digital citizenship within the new onlife public sphere, as well as a strategic node for interpreting and countering the communicative dynamics that fuel online hate speech. However, media education should not be reduced to a merely protective or defensive approach. Rather, it must be conceived as an empowering educational process aimed at developing media competence (Medienkompetenz; Baacke, 1997), enabling individuals to engage with media environments in critical, informed and creative ways (Perfetti, 2020).

In light of the globalisation of communication and the centrality of digital technologies, educational institutions are increasingly called upon to foster what has been described as a 'fourth dimension' of learning: a set of competences related to the critical and ethical use of information and communication technologies. Within this framework, media education can be understood as a multidisciplinary and transversal approach oriented towards the development of critical autonomy (Masterman, 1985) and lifelong media competence (Martínez-de-Toda, 1998). We may therefore argue that media education is closely linked with citizenship education, as a privileged space for mediating participation and fostering practices of reflexivity, responsibility and active engagement oriented towards the construction of a more inclusive and democratic media culture (Adamoli, 2020).

While existing research has mainly focused on perpetrators and victims of hate speech, less attention has been given to those who witness these dynamics without being directly involved. This article addresses this gap by focusing on 'digital spectatorship', often characterised by passive observation within platform-mediated interactions. It argues that spectators play a key role in either reinforcing or challenging online hate speech, thus contributing to debates on the bystander effect, platformed witnessing and digital citizenship. By plac-

¹ According to Buckingham's definition (2003), media education can be understood as the process of teaching and learning *about* the media,

while media literacy represents its outcome in terms of knowledge, skills and critical attitudes.

ing spectators at the centre, the article highlights the link between everyday online practices and broader educational and civic implications.

2. EMPIRICAL PROJECT²

The empirical research is situated within the PRIN 2022 project “Theories, Philosophies and Politics of Bodies in Network-Connected Digital Universes” (Prot. 2022FENML2), which investigates transformations in the relationships between texts, bodies and communicative practices within networked digital universes³. These environments are understood as genuine informal educational contexts and online spaces of socialisation. The research considers digital media not merely as channels of transmission or spaces of expression, but as symbolic environments and socio-technical devices that actively contribute to the construction of meanings, identities and power relations (Meyrowitz, 1985), thereby influencing processes of learning, participation and citizenship.

Drawing on an interdisciplinary approach that integrates political and linguistic philosophy, social psychology, cultural anthropology and the sociology of communication, the project aims to critically rethink the notions of text and body considering digital interactions. It examines how online representations of bodies and discursive practices shape forms of visibility, recognition and exclusion. Within this framework, manifestations of online hate speech are treated as an emblematic phenomenon of the educational tensions that cross the digital public sphere, and as a privileged field for observing both the limits and the potential of individuals’ media competence.

Within this wider research context, the present article specifically focuses on digital spectatorship of online hate speech, with the aim of understanding how individuals who witness discriminatory content interpret and respond to it, and how these processes relate to media competence and educational practices. The epistemological ambition of the study therefore lies in

the co-construction of a critical understanding of the ways in which representations and performativity of bodies in digital spaces affect citizens’ lives, by relating these dynamics to processes of media literacy, empowerment and informed participation. From this perspective, empirical research plays a central role in translating the theoretical framework of media education into empirically grounded evidence capable of informing not only academic debate, but also the design of educational interventions and the development of public policies aimed at promoting a more inclusive and democratic media culture.

The empirical investigation adopted a realist evaluation framework (Pawson & Tilley, 1997; Pawson, 2006), a methodological approach based on empirical evidence that allows the ‘black box’ of the object of study to be opened by analysing the mechanisms underlying the interaction between context and outcomes. This methodology operates through the model M (Mechanism) + C (Context) = O (Outcome) (Pawson & Tilley, 1997, p. 130), framing the inquiry not as a mere description of results, but as an investigation of the processes that generate them, within a constant process of reciprocal causation. Contextual factors were operationalised through socio-demographic variables and patterns of platform use; mechanisms were examined through users’ perceptions, interpretative processes and attitudes (e.g., moral disengagement); outcomes were identified in observed behaviours, such as exposure to hate speech and forms of reaction or non-intervention. The research is therefore structured into two integrated phases: an Evaluative Survey Analysis phase and a phase based on the application of the Nominal Group Technique, aimed at the critical interpretation of the findings through expert judgement⁴.

3. THE EVALUATIVE SURVEY

3.1 Design and analytical orientation

The empirical investigation included the administration of an evaluative survey (Palumbo & Scardigno, 2024) based on a multi-method data collection strategy. The instrument was developed from exploratory research questions focusing on the representation of bodies on social media in relation to normative discourses on gender and sex, on social reactions to bod-

² The authors would like to thank the research team that contributed to the analysis of the relevant literature and to the development of the survey: Maria Rosaria Vitale, postdoctoral research fellow at the University of Urbino; Alessia Franco and Edoardo Maria Bianchi, postdoctoral research fellows at the University of Bari. The authors also wish to thank Teresa Ester Cicirelli and Domenico Emiliano, PhD candidates at the University of Bari, for their significant contribution to the various phases of the research, from the construction of the survey to the analysis of the collected data and the implementation of the Nominal Group Technique phase.

³ The PRIN 2022 project *Theories, Philosophies and Politics of Bodies in Network-Connected Digital Universes* (Prot. 2022FENML2) is coordinated by Professor Francesca Romana Recchia Luciani.

⁴ The data generated through the analysis of the survey and the application of the Nominal Group Technique are intended to be included in a collective publication within the proceedings of the PRIN 2022 project “Theories, Philosophies and Politics of Bodies in Network-Connected Digital Universes” (Prot. 2022FENML2), which is currently in preparation.

ies that deviate from dominant norms and on the ways in which such bodies are excluded or marginalised within the digital ecosystem. These research questions were operationalised through a structured process of conceptual decomposition, in line with a Lazarsfeldian approach, which translated key constructs into analytically manageable dimensions, variables and empirical indicators. The structure of the questionnaire was organised around four main analytical units – bodies and gender, bodies and intersectionality, hate speech and communicative functions, and moderation – which guided the formulation of items. Attention was devoted to examining the frequency and forms of discrimination, verbal violence and hate incitement, as well as users' perceptions of hateful content and of the mechanisms through which such phenomena are normalised, contested or amplified within platforms.

The final questionnaire consisted of 59 items, including cardinal, ordinal and categorical variables and incorporated the use of Likert scales to capture attitudes and perceptions. While no fully validated scales were adopted, item construction was developed through a collaborative team-based process, drawing inspiration from established instruments such as the Moral Disengagement Scale (Bandura et al., 1996), the Coping with Cyberbullying Questionnaire (CWCBQ) and the Hate Speech Intensity Scale⁵. A pre-test was conducted with researchers and participants external to the survey design process, allowing for the refinement of item clarity, coherence and relevance.

In addition to the analysis of quantitative aspects, the survey included a qualitative examination of participants' comments and open-ended responses, thus enabling a complex reading of the phenomena under investigation. The collected data constituted the basis on which the subsequent Nominal Group Technique phase was developed, involving experts in the critical interpretation of perceived differences across target groups. The empirical evidence was analysed not only as descriptive indicators of online communicative practices, but also as interpretative resources for interrogating levels of awareness, responsibility and media competence among the subjects involved. The investigation therefore aims to understand to what extent the ability to recognise, interpret and problematise hateful discourses and discriminatory representations of bodies may inform educational practices oriented towards the development of critical autonomy, communicative responsibility and forms of active and informed digital citizenship. In this sense, the survey does not merely record attitudes and perceptions,

but also contributes to identifying educational nodes and critical issues on which to intervene to promote a more inclusive and reflective media culture.

3.2 Sampling

As the primary target consisted of Italian social media users, a mixed non-probabilistic sampling strategy was adopted, combining two complementary approaches: convenience sampling, based on the voluntary participation of individuals who chose to complete the questionnaire and snowball sampling, whereby respondents were encouraged to disseminate the questionnaire to other potential participants. Given the nature of the recruitment process, the resulting sample cannot be considered representative of the Italian population of social media users and should be interpreted as a situated and context-specific group.

The survey was administered online via the Google Forms platform. Data collection took place between 16 July and 16 October 2025. The instrument was initially distributed through social media channels and personal contacts and subsequently circulated through participants' networks. This mode of recruitment may have influenced the composition of the sample, contributing to its geographical concentration and relatively high educational level, as well as potentially involving individuals with greater sensitivity or interest in the topic. The only inclusion criterion for participation was residence within the Italian territory.

a. Methodology

Data analysis was conducted using an integrated quantitative-qualitative approach. Data preparation, initial cleaning, management of missing data and variable recoding were carried out in Microsoft Excel. Statistical analyses were performed using the Jamovi software, while open-ended responses were recoded and analysed using NVivo.

The research followed a two-stage methodological design implemented through Jamovi. The univariate phase involved the examination of frequency distributions of the variables considered, producing descriptive tables that outlined the aggregated profile of the sample through absolute frequencies and relative percentages. This approach made it possible to assess the overall composition of the studied population and the prevalence of the experiences and opinions investigated. The bivariate phase was conducted through the analysis of contingency tables, which revealed methodologically relevant interrelationships between pairs of variables and identified key structural intersections within the

⁵ The latter were used only in specific sections of the broader survey not directly addressed in this article.

sample. The use of chi-square (χ^2) tests complemented the descriptive analysis by assessing the statistical significance of the observed associations, while association indices (Phi, Cramér's V) were employed to quantify the strength of these relationships.

Missing data were handled through a case-sensitive approach. Relative frequencies were calculated based on valid responses for each item, in order to avoid distortions due to non-response. For inferential analyses, cases with missing values were excluded where necessary, ensuring the robustness of statistical testing without introducing artificial estimations.

Variable recoding was applied selectively to enhance analytical clarity and ensure statistical reliability. In particular, the geographical variable was recoded into a dichotomous form (Puglia vs. the rest of Italy) due to the marked regional imbalance within the sample. Similarly, items related to moral disengagement were grouped into broader conceptual dimensions drawing on Bandura's theoretical framework. These clusters were then binary recoded (presence vs. absence of the attitude), enabling the construction of descriptive frequency tables and facilitating cross-tabulation.

Open-ended responses were analysed through a structured qualitative coding process aimed at identifying recurring patterns of interpretation and behaviour. Data were imported into NVivo and coded using a bottom-up approach to identify recurrent behavioural categories. The coding process was conducted by a single coder, and no intercoder reliability procedures were applied. Nonetheless, to enhance internal coherence, iterative cycles of code refinement were employed, ensuring

consistency across categories (reporting, defensive intervention, non-intervention).

b. Structural analysis of the sample

The empirical investigation produced a final dataset of $N = 457$ valid cases. The analysis of the structural variables reveals a sample affected by specific selection biases, which delimit its scope of representativeness. From a gender perspective, there is a marked predominance of women (63.0%) compared to men (35.5%). The age distribution shows a strong concentration in adulthood, with the 26-35 age group as the statistical mode (43.4%), while the 36-50 and 51-65 age groups are both consistently represented (13.0% each). Although a minority, the over-65 group is not negligible (6.2%). Overall, the age profile confirms the extension of digitalisation processes to more mature cohorts (Fig. 1).

From a territorial point of view, the sample includes respondents from 14 Italian regions but displays a strong over-representation of Puglia ($N = 325$, 75.1% of the total, Fig. 2), in line with the adopted sampling strategy. The low frequencies observed in the other regions do not allow for robust statistical comparisons. Consequently, the geographical variable was recoded in a dichotomous form (Puglia vs the rest of Italy) in order to ensure greater analytical reliability in subsequent analyses. At the same time, this territorial concentration should be taken into account when interpreting the findings, as it may reflect context-specific cultural and social dynamics that influence perceptions and responses to hate speech.

The sample is also characterised by a high level of cultural capital, with a pronounced high-education

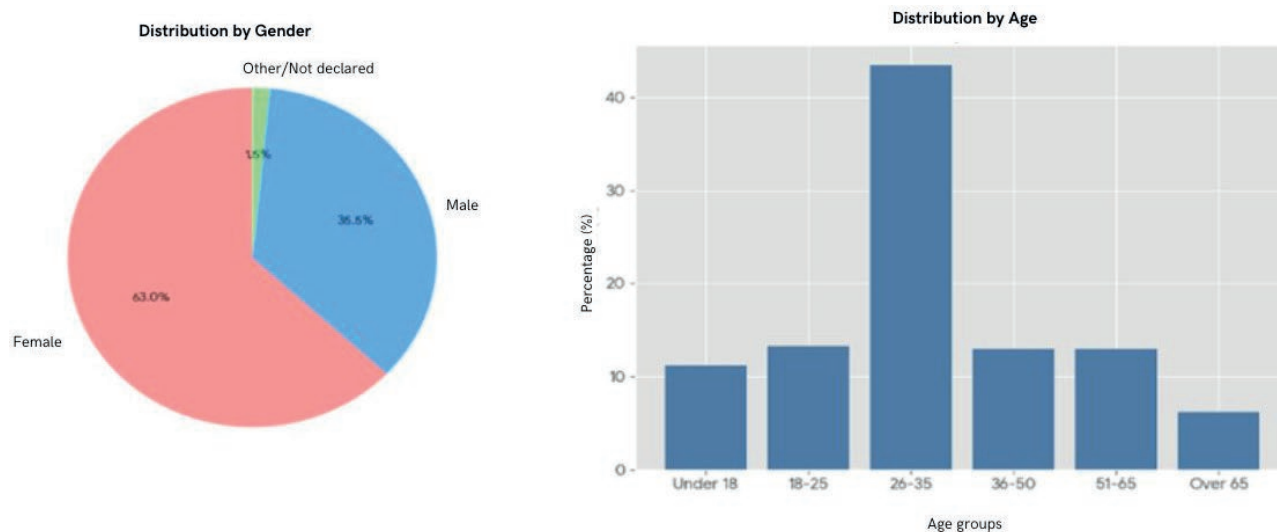


Figure 1. Distribution of the sample by gender and age. The charts and tables were produced using the Jamovi software or Microsoft Excel.

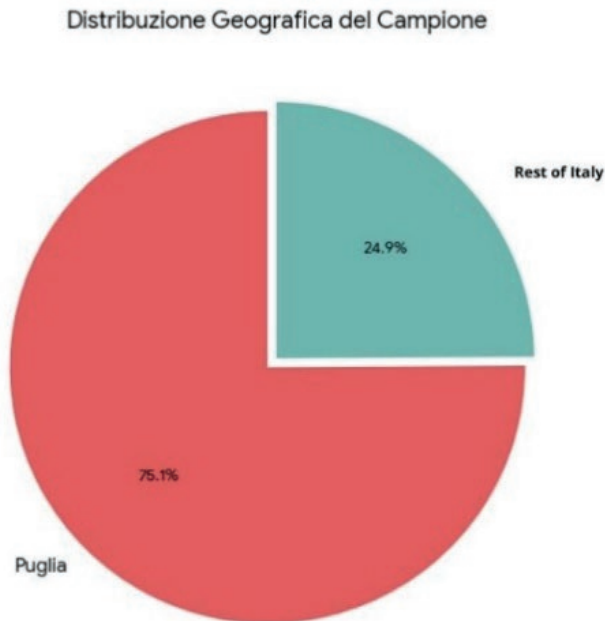


Figure 2. Geographical distribution of the sample.

bias: 63.5% of respondents hold a university or post-graduate qualification. This profile is reflected in the occupational composition, which includes a substantial proportion of students (27.4%), including doctoral candidates, alongside employees and self-employed professionals. Such a composition may be associated with higher levels of awareness and interpretative sensitivity towards discriminatory discourse, potentially influencing both the recognition of hate speech and the propensity to reflect on or respond to it. Finally, patterns of digital consumption reveal a strong platform- and age-based polarisation, outlining a generational segmentation of digital spaces: Instagram is predominant among young adults, TikTok among Generation Z and Facebook among older cohorts, with no evident gender-based differences (Fig. 3).

4. WITNESSING ONLINE HATE SPEECH: DYNAMICS, PERCEPTIONS AND BEHAVIOURS

The data collected through the survey were organised into three main thematic areas: representations of online hate speech, perceptions of online hate speech and moderation practices. This paper focuses on social media users who have witnessed episodes of online hate speech without being direct victims, with the aim of describing the observed dynamics and the behaviours

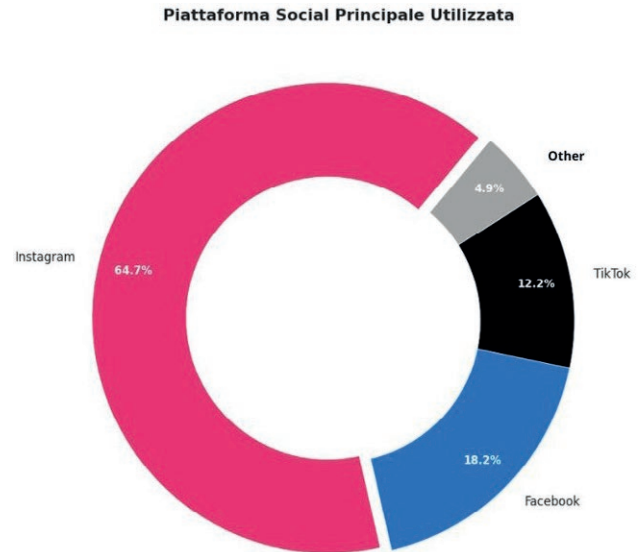


Figure 3. Distribution of social media use.

that result from them. The analysis was guided by the following research questions⁶:

- Q1: How many participants have witnessed episodes of online hate speech, and how frequently?
- Q2: What is the main underlying dimension of the online hateful content observed?
- Q3: How do individuals who witness hate speech react?
- Q4: Do social media users display attitudes of moral disengagement?

Q1: How many participants have witnessed episodes of online hate speech, and how frequently?

The survey analysis shows that most participants have witnessed episodes of online hate speech with some

⁶ Methodological note: in order to conduct a targeted analysis of the dynamics related to the experience of online hate speech, a filter was applied to the overall dataset to isolate only those participants who reported having observed at least one episode of hate speech on social media. This step made it possible to focus the analysis on a subsample representing 89% of the total sample, which contained the data necessary to address the research questions. In the research questionnaire, responses to the questions on the frequency of exposure to hate speech (specifically 'Rarely', 'Sometimes' and 'Often') triggered a series of additional drop-down questions designed to further explore the types of episodes observed and the subsequent behavioural reactions. This branching structure made the application of the filter not only appropriate but methodologically necessary to correctly analyse the sections of the questionnaire that were available exclusively to respondents with experience of exposure to hate speech. By contrast, the dimensions related to moral disengagement were measured through a set of Likert-scale items administered to the entire sample.

degree of frequency (Tab. 1). Only 10.3% of respondents report having never encountered such episodes, while 15.8% state that they have observed them rarely. A substantial proportion of participants (37.4%) report having witnessed hate speech ‘sometimes’ (2–4 episodes), and 36.1% indicate that they have seen it ‘often’ (five or more episodes). Only a very small share of respondents (0.4%) report that they do not know or cannot recall the frequency. Overall, these data indicate that 89% of the sample – 409 respondents out of 457 – have witnessed at least one episode of digital hate speech. Approximately 73.5% of participants report repeated exposure to online hate speech (from ‘sometimes’ upwards), highlighting both the pervasiveness and the consistency of the phenomenon within the analysed population.

Table 1. Frequency of observation.

FR_Frequency observation	Frequencies	% del Total	% Cumulative
0 (never)	47	10.3%	10.3%
1 (rarely)	72	15.8%	26.0%
3 (sometimes)	171	37.4%	63.5%
5 (often)	165	36.1%	99.6%
9 (Don't know / cannot recall)	2	0.4%	100.0%

The analysis of the relationship between demographic variables and the frequency of witnessing online hate speech produced differentiated results by gender and age. While no statistically significant association emerged between gender and frequency of exposure, age appears to be a significant factor in shaping patterns of exposure (Fig. 4).

Test χ^2			
	Valore	gdl	P
χ^2	8.10	16	0.946
N	457		

The chi-square test ($\chi^2 = 65.3$; $p < 0.001$) confirms the existence of a statistically significant relationship between age groups and the frequency with which hate speech episodes are observed. Central age groups (2: 18–25, 3: 26–35, and 4: 36–50) report higher levels of exposure or awareness of the phenomenon.

Test χ^2			
	Valore	gdl	P
χ^2	65.3	24	<.001
N	457		

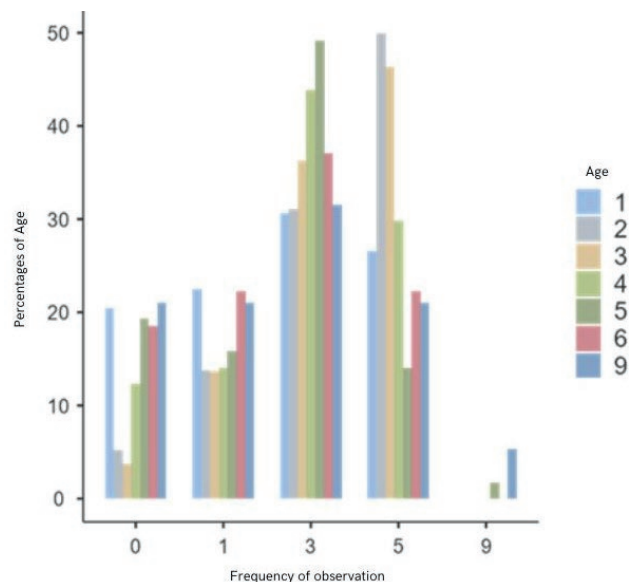


Figure 4. Relationship between frequency of observation and age groups.

In particular, 50.0% of participants in age group 2 report having observed hate speech ‘often’ (five or more episodes), while the corresponding proportions are 46.3% for group 3 and 29.8% for group 4. By contrast, participants in the youngest (1: under 18) and oldest age groups (5: 51–65 and 6: over 65) tend to report less frequent observations, with higher percentages in the ‘never’ or ‘rarely’ categories⁷.

The analysis also considered the relationship between having been a victim of hate speech and having witnessed at least one episode of digital hate speech. Within the total sample ($N = 457$), 32% of participants report having personally experienced hate speech on social media. The contingency table analysis reveals a significant association between having been targeted by hate speech (RO – receipt of offences) and the frequency with which users report witnessing hostile behaviour directed at others (FR – frequency of observation). The chi-square test confirms that these two variables are not independent.

Test χ^2			
	Valore	gdl	p
χ^2	19.4	8	0.013
N	457		

⁷ In order to mitigate the bias associated with the mixed non-probabilistic sampling strategy, which combines convenience and snowball sampling, exposure rates to hate speech were calculated within each age subgroup. This approach enabled a more differentiated and robust analysis of exposure patterns across the different demographic segments.

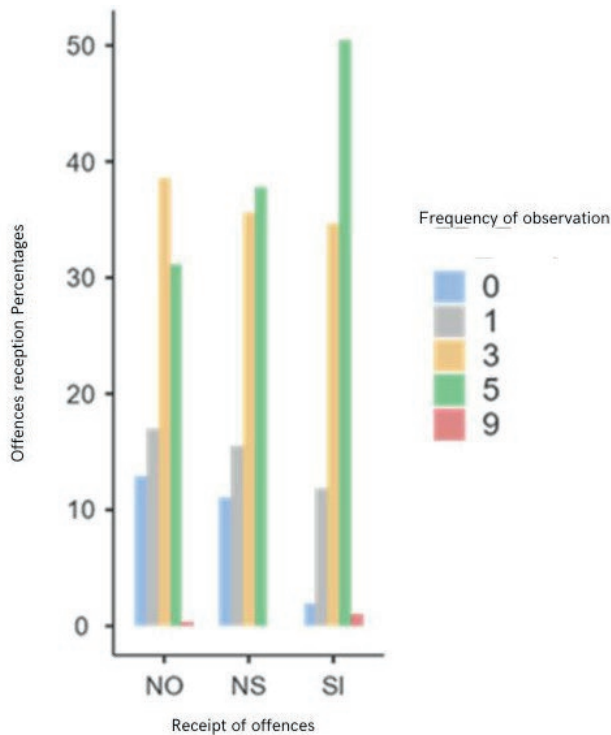


Figure 5. Relationship between witnessing hate speech and having experienced it.

Users who report having been victims of hate speech tend to observe such episodes more frequently than those who have not experienced direct victimisation (Fig. 5). This pattern may be interpreted in different, not mutually exclusive ways. On the one hand, individuals who have been targeted may be more frequently exposed to contexts in which such behaviours occur; on the other, prior experience of victimisation may be associated with a greater attentiveness to or awareness of hostile interactions (Waldron, 2012; Goffman, 1963). While the present data do not allow these dimensions to be clearly disentangled, they suggest that personal experience of hate speech is linked to differences in how such phenomena are perceived and reported.

Q2: What is the main underlying dimension of the online hateful content observed?

The in-depth analysis of the survey results indicates that the primary target dimensions of online hateful content relate to physical appearance, sexual orientation and political opinions. These are followed, with lower but still significant frequencies, by dimensions related to ethnicity and gender (Fig. 6). The prominence of physical appearance as a primary target

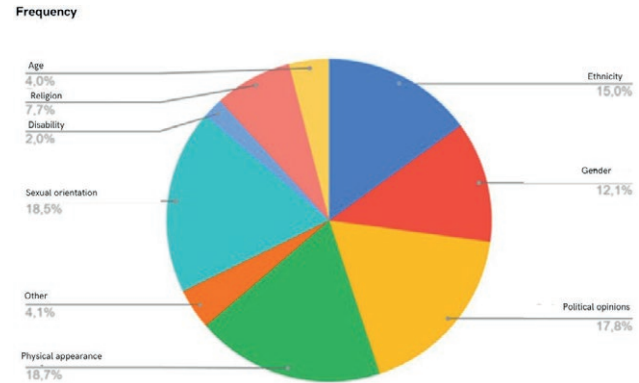


Figure 6. Target dimensions of hate speech.

appears particularly relevant considering the centrality attributed to the body within the theoretical framework of this study. In digital environments characterised by high levels of visibility and rapid circulation of images and representations, the body may become a key site of exposure and evaluation. Normative expectations concerning what bodies should look like, often shaped by dominant aesthetic, gendered and cultural standards, can be reproduced and reinforced through online interactions. Within this context, bodies that deviate from these norms may be more easily perceived as ‘other’ and, as such, more frequently exposed to hostile or exclusionary discourse.

Q3: How do individuals who witness hate speech react?

To investigate the reactions of individuals who have witnessed episodes of online hate speech, the survey included an open-ended question. The use of this unstructured item made it possible to collect detailed descriptions of reactions and actions taken, capturing nuances and behavioural complexity that would have been difficult to detect through closed-ended items. Responses were subsequently subjected to systematic qualitative coding: each response was carefully read and categorised according to the type of reaction reported (identified categories: reporting, defensive intervention, and no intervention). Attention was devoted to the treatment of ambiguous or mixed responses. In cases where participants described multiple actions, the dominant or most explicitly emphasised behaviour was selected for coding, while responses that were difficult to classify were discussed in relation to their contextual meaning and assigned to the most coherent category. The final set of categories should therefore be understood as the result of a process of analytical synthesis, aimed at balanc-

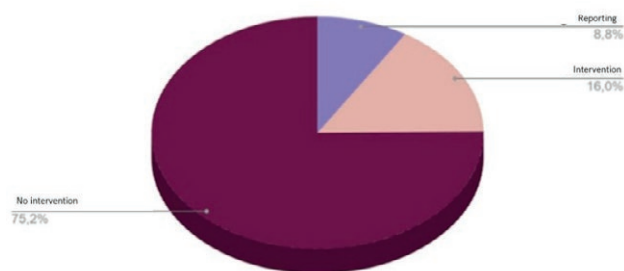


Figure 7. User reactions to online hate speech.

ing descriptive richness with the need for comparability across cases. The coding process was conducted using NVivo, which enabled the management and aggregation of textual data, the synthesis of emergent categories and the quantification of their distribution within the subsample. This approach transformed qualitative responses into statistically analysable categorical variables, while preserving the informational richness of participants' reported experiences.

The survey results show that, when faced with episodes of hate speech, most individuals tend not to intervene actively: 75% of respondents report that they did not take part in the discussion. A smaller proportion (16%) state that they intervened by commenting or defending the targeted individual, while only 9% chose to report the discriminatory content (Fig. 7). Overall, the data reveal a predominance of passive attitudes and limited engagement in counteractive actions. This pattern reflects the well-documented «bystander effect» (Latané & Darley, 1970; Markey, 2000), namely the tendency of observers not to intervene in critical situations, which appears particularly pronounced in digital environments.

This tendency can also be interpreted in light of Computer-Mediated Communication models. According to the Social Identity Deindividuation (SIDE) model (Spears & Lea, 1992), when contextual cues such as appearance, status or age are reduced, individuals tend to perceive themselves less as unique subjects and more as members of a social group. Drawing on Social Identity Theory (Tajfel, 1978) and Self-Categorization Theory (Turner et al., 1987), this perspective suggests that anonymity does not simply foster disinhibited or antisocial behaviour, but may also strengthen alignment with group norms when social identity becomes salient. In highly populated digital environments, this may reinforce passive conformity when non-intervention appears to be the dominant behaviour.

Q4: Do social media users display attitudes of moral disengagement?

The analysis indicates that social media users do indeed tend to display attitudes due to mechanisms of moral disengagement, as theorised by Bandura and operationalised in the Moral Disengagement Scale (1996), which informed the construction of the items used in this survey⁸. These processes appear to be amplified in digital contexts, where the mass presence of spectators and technological mediation foster dynamics of collective de-responsibilisation⁹. In this sense, the reduced visibility of individual identity and the salience of group dynamics in online environments may favour processes of deindividuation, as suggested by the SIDE model, thus contributing to the diffusion of responsibility observed in the data. Moreover, platform architectures contribute to the redistribution and obscuring of moderation responsibilities, encouraging a systemic delegation that reduces users' perception of individual agency (Gillespie, 2018).

In particular, the results reveal a marked prevalence of two mechanisms (Fig. 8): diffusion of responsibility

⁸ In constructing the items related to moral disengagement, reference was made to Bandura's Moral Disengagement Scale (1996), adopting a revised version adapted to the context of digital communication. Bandura identifies eight mechanisms of moral disengagement: moral justification, advantageous comparison, euphemistic language, displacement of responsibility, diffusion of responsibility, minimisation or distortion of consequences, attribution of blame and dehumanisation of the victim. For reasons of effectiveness and brevity in online administration, as well as to emphasise the processes most relevant in digital contexts, a reduced version comprising six conceptual clusters was adopted. Accordingly, two clusters from the original scale were merged with conceptually related ones in order to reduce item redundancy and enhance the interpretative clarity of responses. Specifically, the mechanism of displacement of responsibility was integrated into the items related to diffusion of responsibility, as both can be traced back to processes of de-responsibilisation of individual agency, which are amplified in digital environments characterised by a real or perceived plurality of observers. Similarly, the mechanism of minimisation or distortion of consequences was incorporated into the items relating to both euphemistic language and advantageous comparison. This choice is motivated by their shared function of attenuating the perceived severity of hostile conduct: in the former case through linguistic and semantic strategies that reduce the emotional impact of harm and in the latter through comparative processes that relativise the action in relation to behaviours perceived as more serious. In digital contexts, these mechanisms tend to co-occur and overlap in discursive practices, making a clear empirical distinction between them difficult. This reorganisation made it possible to ensure theoretically adequate coverage of the main mechanisms of moral disengagement, while maintaining a more streamlined measurement instrument, consistent with the communicative specificities of social media and functional to the analysis of passivity and non-intervention in the presence of online hate speech.

⁹ These dynamics can be linked to the bystander effect described by Latané and Darley (1970) and subsequently adapted to computer-mediated communication environments by Markey (2000).

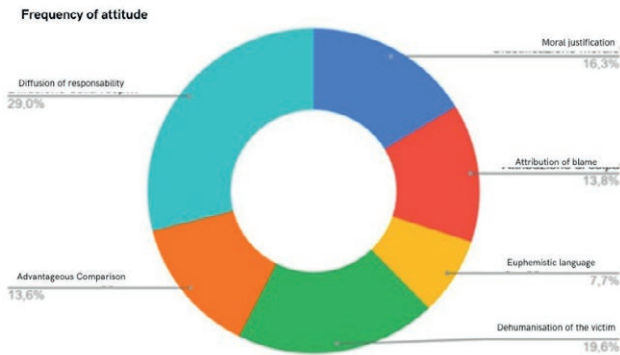


Figure 8. Attitudes of moral disengagement.

and dehumanisation of the victim. Diffusion of responsibility emerges as the most widespread process, with 29% of users acknowledging this attitude. This finding supports the theory that the real or perceived presence of numerous online observers reduces individual responsibility, fostering a tendency towards inaction. In digital environments, the massive and often anonymous nature of social media significantly amplifies this phenomenon, leading to systemic de-responsibilisation. Dehumanisation of the victim represents the second most frequent mechanism, involving 20% of the sample. This cognitive and relational process allows individuals to perceive others as lacking full humanity, thereby reducing empathy and emotional resonance in the face of hostile or conflictual content. Such distancing facilitates psychological disengagement and a lower propensity to intervene¹⁰. Moreover, the affordances of computer-mediated communication allow for selective self-presentation and limited exposure to contextual cues, which, as highlighted by the Hyperpersonal model (Walther, 1996), may intensify interaction processes. In such environments, users can strategically control the information they disclose, idealise their interlocutors on the basis of partial cues and construct messages more reflectively, thereby amplifying both relational involvement and the polarisation of evaluative judgments.

¹⁰ It should nevertheless be noted that the reliability of responses to these items may be partially compromised by social desirability bias. As these attitudes and cognitive processes are socially stigmatised, it is plausible that participants tended to underestimate their own involvement in mechanisms of moral disengagement, providing responses perceived as more socially acceptable. This bias represents an inherent limitation of self-reported measures of moral constructs and should be considered when interpreting the results.

5. NOMINAL GROUP TECHNIQUE

The subsequent phase of the research applied an online and asynchronous adaptation of the Nominal Group Technique (NGT) (Delbecq & Van de Ven, 1971; Duncan Smith et al., 2024; Dal Mas et al., 2023a e 2023b) in order to involve experts in a critical evaluation of the survey findings and in the formulation of operational strategies focused on the theme of the 'body'. The process was structured in three phases:

1. preliminary individual interviews aimed at collecting exploratory contributions;
2. qualitative analysis and thematic coding of the materials, conducted using NVivo software, with the aim of constructing a structured set of operational items;
3. an asynchronous evaluation and prioritisation phase, in which experts expressed their judgements on the proposed items.

Six experts participated in the interviews, selected on the basis of criteria of high theoretical and practical expertise in the field of online hate speech. In particular, the selection considered: academic expertise in areas such as gender studies, disability and queer studies, critical pedagogy and inclusive communication; professional or activist experience in transfeminism, queer, crip or LGBTQIA+ contexts; and positions of responsibility in research, public engagement, cultural production or the DEI sector.

Regarding the macro-theme of the perception of online hate speech, the question submitted to the experts was the following:

Based on the survey results, it emerges that online discrimination phenomena are widespread (73% of users observe discriminatory content sometimes or often) and that counteractive interventions remain limited: only 16% comment or defend victims, 9% report the content, while 75% do not intervene at all. Moreover, certain attitudes such as diffusion of responsibility and dehumanisation of the victim appear to be relatively widespread and may represent significant psychological barriers to active engagement. Considering these data, which concrete interventions do you believe could be implemented to raise awareness and transform the passive perception of hate into active counteraction and prosocial interventions, particularly when the target concerns individuals whose bodies or identities diverge from dominant normative representations?

The interviews, once transcribed and coded, generated operational strategies clustered into macro-areas, which were then submitted to the NGT to prioritise mechanisms for countering hate speech and addressing

the emotional vulnerability of victims. Each expert had access exclusively to the list of items¹¹ to be evaluated and assigned priorities individually and independently, without any possibility of mutual influence. Evaluations were collected anonymously, in order to protect participants' confidentiality and the integrity of the process. This procedure enabled straightforward score extraction and accurate data analysis. The outcomes of the evaluative process will ultimately feed into a dissemination output, culminating in the production of a *vademecum* dedicated to the valorisation of the theme of the 'body' and oriented towards the definition of guidelines addressed to public decision-makers.

5.1 Results of the NGT

In the evaluation phase of the NGT method (Bertin, 2011), experts' contributions were analysed with respect to the effectiveness and priority of operational strategies for countering online hate speech. The NGT results show a strong convergence among experts towards an integrated and multi-level model for addressing online hate speech, which articulates interdependent interventions at the individual, community and educational-cultural levels, with significant differences in the prioritisation of specific actions (Bronfenbrenner, 2005).

At the level of individual intervention, an interpretative tension emerges between a perspective that prioritises the recognition of hate speech as a cognitive and cultural prerequisite for action and an orientation more strongly focused on immediate relational intervention. While one expert assigns the highest priority to recognising subtle and normalised forms of hostility, other experts place direct support for the victim at the top, understood as an action capable of immediately affecting dynamics of isolation and symbolic rebalancing in digital spaces. Reporting hostile content, although recognised by all as a necessary and accessible practice, tends to occupy a subordinate or intermediate position, signalling a shared assessment of its structural limitations when not accompanied by broader processes of awareness and responsibility-taking. At the level of coordinated community intervention, a stronger convergence is observed. Self-organised community moderation and collective support actions emerge as the strategies with the greatest potential impact, as they can amplify responses, interrupt the spiral of silence and alter perceptions of social consensus in hostile contexts. The interruption of the spiral of silence is frequently

positioned immediately afterwards, as a complementary practice that strengthens alliances and reduces the intimidating effect of hate speech. By contrast, the re-signification of hostile terms, although recognised as culturally relevant, is consistently placed last, likely due to its greater complexity and the risk of ambiguous outcomes in the absence of strong community support. Within the educational-cultural domain, evaluations display both convergence and differentiation. School-based education and widespread training are unanimously recognised as fundamental strategic levers for structural and lasting change. Differences emerge, however, in the assessment of support actions: some experts assign greater relevance to everyday micro-interventions and 'sentinel' figures as mechanisms for lowering the threshold for activation, while others place greater value on public awareness campaigns and reflexivity devices at the point of platform access. Overall, these differences do not constitute conflicting positions but rather reflect situated and contextual readings of the effectiveness of strategies, reinforcing the interpretation of the proposed actions as complementary and flexibly activatable depending on resources, contexts and the subjective conditions of those involved.

Taken together, the evaluations delineate an orientation that privileges diffuse responsibility over exclusive delegation to platforms and conceives the countering of hate speech as a continuous, situated and relational educational process. This interpretation is consistent with perspectives from Adaptive Structuration Theory (DeSanctis & Poole, 1994), according to which users actively appropriate technological environments, reproducing or transforming communicative norms through their situated practices. Within this framework, the critical position expressed by one expert regarding the possibility of rigid hierarchisation of strategies further reinforces the interpretation of the proposed interventions as complementary and context-dependent, shaped by available resources and subjective conditions (Gagliardone et al., 2015).

6. EDUCATIONAL IMPLICATIONS

The findings emerging from the empirical research strongly confirm the urgency of situating the countering of hate speech and discriminatory representations of bodies within a broader educational framework¹². The evidence collected shows that online communicative practices are permeated by dynamics of normalisation

¹¹ The list of the operational strategies identified can be consulted in the appendix.

¹² *Digital Competence Framework for Citizens* (DigComp 2.1).

of symbolic violence, stereotyping and exclusion, which can hardly be addressed through purely normative or repressive interventions. Rather, they call for the development of critical, reflective and ethically oriented competences, capable of acting upon the processes of production, circulation and interpretation of media content. These dynamics can also be understood as situated and interactional processes, in line with Situated Action Theory (Suchman, 1987), in which meaning and behaviour emerge within specific communicative contexts rather than from predefined intentions. From this perspective, the empirical data prove particularly prolific when read in light of the construction of media competence, which identifies media education as a pathway aimed not only at understanding media languages and strategies, but also at developing awareness of the economic, political and cultural logics that shape media functioning, as well as the capacity to assume an active and responsible role in encoding and decoding processes. Educational interventions should therefore aim to develop critical media literacy skills oriented towards recognising explicit and implicit forms of hate speech, processes of normalisation and algorithmic logics that amplify visibility, alongside competences in production and participation that allow individuals to experiment with alternative, inclusive and solidaristic communicative practices. This entails education in the conscious use of reporting practices, public and private support for victims and community moderation. The findings concerning high exposure to hateful content, the prevalence of non-intervention and the diffusion of moral disengagement mechanisms indicate a widespread fragility not so much in technical skills as in critical, social and ethical competences of digital action.

In concrete terms, this may translate into structured school-based modules combining analytical and experiential activities, such as guided exercises in the recognition of hate speech through real or simulated online interactions, the critical analysis of platform algorithms and visibility mechanisms (platform literacy) and workshops in which students collaboratively redesign harmful content into inclusive narratives. Peer education practices can further reinforce these processes, for instance through student-led discussions or “digital ambassadors” programmes aimed at fostering responsibility within peer groups. At the same time, specific intervention protocols for spectators can be introduced, training students in graduated forms of response (e.g. reporting, public counter-speech, private support to victims), as well as in the conscious use of micro-actions such as signalling, commenting or expressing dissent in non-escalatory ways. Such approaches may also include role-playing

activities that simulate bystander situations, helping learners to rehearse ethical decision-making in context.

In a context in which young people are no longer merely consumers but also producers of content (Rivoltella & Ardizzone, 2007), the principle of responsibility (Perfetti, 2020) becomes especially salient, calling education to intervene in choices before they translate into harmful or exclusionary communicative practices. In this direction, the research results suggest the value of designing media education interventions that integrate cognitive, ethical-value-based and emotional dimensions. The mechanisms of dehumanisation and diffusion of responsibility identified in the empirical analysis highlight how a lack of empathy and emotional resonance constitutes a barrier to active intervention. The spread of hateful language and verbal violence cannot be understood, nor effectively countered, without education in emotions and the acquisition of socio-emotional competences, which are essential for fostering empathy, proximity and the ability to recognise others as legitimate interlocutors (Bornatici, 2017). In this perspective, schools emerge as a crucial educational space for the construction of a media culture oriented towards the common good, in line with the principles of the 2030 Agenda, the Education for Sustainability Plan and the Digital Civic Education Curriculum, which explicitly recognise the role of media education as a space for the construction of *educational commons* (Cappello, 2010), where competences, values and practices are collectively shared and negotiated. Countering hate speech ultimately means investing in a pedagogy of the ‘us’ (Bruzzzone, 2016), capable of valuing differences, promoting active participation and supporting the construction of critical and responsible subjectivities, able to inhabit media spaces as arenas of relationship, dialogue and democratic citizenship.

7. CONCLUSIONS

The empirical evidence discussed in this paper confirms that online hate speech cannot be interpreted solely as a set of deviant individual acts, but must be understood as a structural phenomenon of digital communicative ecosystems, in which platform architectures, implicit social norms and forms of spectatorship, often characterised by passivity and de-responsibilisation, interact. Within this framework, the figure of the bystander assumes an ambivalent role: while repeated exposure to hostile content tends to produce habituation, silence or emotional distancing, it also represents a potential space for ethical and civic activation, provided that adequate

critical competences, emotional resources and educational devices are in place.

The research findings nevertheless show that, in most cases, reactions to hate speech are situated along a continuum ranging from silence to direct intervention, passing through indirect or symbolic forms of response. These modalities should not be read merely as a lack of sensitivity or willingness, but as expressions of structural limits in the formation of media competence, communicative responsibility and emotional regulation in digital environments. In this context, media education is called upon to move beyond a conception focused exclusively on educating the critical spectator, towards the formation of responsible subjects capable of recognising their active role in both the consumption and production of content (Rivoltella & Ardizzone, 2007). From this perspective, educating for digital citizenship means placing the principle of responsibility at the centre, articulated not only at the cognitive level but also at the emotional and value-based levels. Promoting civic sense, empathy and the capacity to recognise others as worthy of respect is possible only if adults – teachers, educators and families – can design educational interventions that engage with the cultural transformations of the onlife condition, rather than merely reproducing pedagogical models of the past (Fabbri, 2008). Choosing how to inhabit the continuous flow of change that characterises digital contemporaneity therefore demands a pedagogical and ethical attitude: designing a relationship with technologies that is sustainable, reflective and oriented towards the common good.

This study contributes to the current debate by shifting the analytical focus from perpetrators and victims to the often-overlooked role of digital spectatorship, thus foregrounding the educational significance of bystander positions and their transformative potential within media education frameworks. At the same time, the findings must be interpreted in light of several limitations, including the non-representative nature of the sample, its geographical concentration and relatively high educational level, as well as the reliance on self-reported data within a cross-sectional design. These constraints call for caution in generalising the results and highlight the need for further empirical investigation. Future research should aim to broaden the empirical base through more diverse and representative samples, adopt longitudinal approaches capable of capturing changes over time and develop experimental or intervention-based designs to assess the effectiveness of targeted media education practices in fostering more active, responsible and critically engaged forms of digital participation.

REFERENCES

- Adamoli, M. (2020). La media education come quadro educativo per un approccio etico alla rivoluzione digitale. *IUSVEducation*, 15, 22-35. <https://www.iusveducation.it/la-media-education-come-quadro-educativo-per-un-approccio-etico-alla-rivoluzione-digitale/>
- Baacke, D. (1997). *Medienpädagogik*. Tübingen.
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, 71(2), 364–374. <https://doi.org/10.1037/0022-3514.71.2.364>
- Bertin, G. (2011). *Con-sensus method*. FrancoAngeli.
- Bornatici, S. (2017). La scuola: Bene comune per educare alla sostenibilità. *MeTis: Mondi educativi. Temi, indagini, suggestioni*, 7(2), 241–248.
- Bronfenbrenner, U. (2005). *Making Human Beings Human: Bioecological Perspectives on Human Development*. Sage.
- Bruzzone, D., (2016). Il sentimento del noi e le sue ombre: provocazioni e suggestioni per educare ad un'identità inclusiva. *Pedagogia oggi*, 2, 130-142.
- Buckingham, D. (2003). *Media Education: Literacy, Learning and Contemporary Culture*. Polity Press.
- Cappello, G.M. (2010). Analisi critica vs. produzione creativa. Le nuove sfide della media education nell'era digitale. *FORM@RE*, 10(70). <https://doi.org/10.13128/formare-12530>
- Dal Mas, F., et al. (2023a). A multidisciplinary approach to surgical care: The case of disease units. A Delphi consensus on the newly-born pancreas units. In J. Martellucci & F. Dal Mas (Eds.), *Towards the future of surgery*, 147–171. Springer. https://doi.org/10.1007/978-3-031-47623-5_13
- Dal Mas, F., et al. (2023b). Towards the healthcare of the future: A Delphi consensus on environmental sustainability issues. In J. Martellucci & F. Dal Mas (Eds.), *Towards the future of surgery: New paradigms in healthcare*, 189-209. Springer. https://doi.org/10.1007/978-3-031-47623-5_15
- Delbecq, A. L. & Van de Ven, A. H. (1971). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science*, 7(4), 466–492. <https://doi.org/10.1177/002188637100700404>
- DeSanctis, G. & Poole, M. S. (1994). Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory. *Organization Science*, 5(2), 121–147. <http://www.jstor.org/stable/2635011>
- Duncan Smith, A., et al. (2024). Use of nominal group technique methods in the virtual setting: A reflective account and recommendations for practice.

- Australian Critical Care*, 37(1), 158–165. <https://doi.org/10.1016/j.aucc.2023.09.004>
- Fabbri, M. (2008). *Problemi d'empatia: La pedagogia delle emozioni di fronte al mutamento*. Etis.
- Floridi, L. (2015). *The onlife manifesto: Being human in a hyperconnected era*. Springer.
- Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000233231>
- Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Penguin Books.
- Goffman, E. (1956). *The presentation of self in everyday life*. Social Sciences Research Centre.
- Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Century Crofts.
- Lazarsfeld, P. F. (1965). Dai concetti agli indici empirici. In R. Boudon & P. F. Lazarsfeld, *L'analisi empirica nelle scienze sociali: Dai concetti agli indici empirici* (Vol. 1, pp. 41–52). Il Mulino.
- Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, 16(2), 183–188. [https://doi.org/10.1016/S0747-5632\(99\)00056-4](https://doi.org/10.1016/S0747-5632(99)00056-4)
- Martínez de Toda, J. (1998). *Metodología evaluativa de la educación para los medios: Su aplicación con un instrumento multidimensional*. Pontificia Universitas Gregoriana, Facultas Scientiarum Socialium.
- Masterman, L. (1985). *Teaching the media*. Comedia.
- Meyrowitz, J. (1985). *Oltre il senso del luogo: L'impatto dei media elettronici sul comportamento sociale*. Baskerville.
- Palumbo, M. & Scardigno, A. F. (2024). Lunga vita alla valutazione (della terza missione) degli atenei: Tra rendicontazione sociale e apprendimento permanente. *Rassegna Italiana di Valutazione*, 88–89, 17–32. <https://doi.org/10.3280/RIV2024-088002>
- Pawson, R. (2006). *Evidence-based policy: A realist perspective*. Sage.
- Pawson, R. & Tilley, N. (1997). *Realistic evaluation*. Sage.
- Perfetti, S. (2020). Il fenomeno dell'hate speech e la cultura digitale: La scuola degli affetti come bene comune. *Encyclopaideia*, 24(56), 119–132. <https://doi.org/10.6092/issn.1825-8670/10749>
- Rivoltella, P. C., & Ardizzone, P. (2007). La media education fra tradizione e sfida del presente. *New Media Education*, 52(15), 50–52.
- Spears, R., & Lea, M. (1992). Social influence and the influence of the 'social' in computer-mediated communication. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 30–65). Harvester-Wheatsheaf.
- Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.
- Tajfel, H. (Ed.). (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations*. Academic Press.
- Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.
- Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.
- Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1), 3–43. <https://doi.org/10.1177/009365096023001001>

APPENDIX

Items Emerging from the Interviews – Operational Strategies to Transform the Passive Perception of Hate Speech into Active Counteraction

1. Individual-level intervention

Reporting: promote the immediate reporting of hostile content (including anonymous reporting) as a minimal action that is always feasible.

Direct support to the victim: send private messages of support to the targeted person (recognition, care and listening), and enhance the victim's visibility through likes, supportive comments and shares.

Recognition of hate speech: train users to identify also subtle forms of hostility (such as dehumanisation and normalised devaluation), and disseminate micro-educational resources (short examples, checklists and linguistic indicators) to facilitate recognition.

2. Coordinated community-level intervention

Community moderation: establish self-organised moderation groups capable of intervening through multiple reports and coordinated responses, and support collective actions in defence of the targeted person (multiple comments, coordinated private messages).

Disrupting the spiral of silence: produce positive, supportive content within hostile threads in order to increase the perception of alliance and counter perceived isolation.

Re-signification of hostile terms: encourage practices of re-signifying stigmatising terms in order to remove them from a negative framing.

3. Educational and cultural intervention

School education and widespread training: integrate continuous modules on the recognition of hate speech into school curricula, and develop training pathways in workplaces and social spaces (community centres, associations and trade unions).

Public awareness campaigns: launch pervasive campaigns across media and public spaces with clear messages highlighting the responsibility to report inappropriate content.

Everyday micro-interventions: encourage small, recurring actions (such as disapproving emojis, brief comments and supportive likes), and promote the presence of 'sentinel' figures in formal and informal spaces who intervene through minimal gestures.

Reflexivity at platform access: require a brief moment of awareness regarding platform values at the point of registration, and introduce short reminders about responsible use when users are about to publish potentially problematic content.