ECHOES OF HATE

DIGITAL COMMUNICATION, POPULISM, AND THE REGULATION OF HATE SPEECH

Eliza Bechtold, Clara Burbano-Herrera,
Zoë Grossi, Yves Haeck (eds.)

Echoes of Hate
Digital Communication, Populism and the Regulation of Hate Speech

edited by
Eliza Bechtold, Clara Burbano-Herrera, Zoë Grossi, Yves Haeck

First published in Belgium, 2025
Human Rights in Context, Ghent University

Second edition: February 2026

ISBN : *9789076566009*

Attribution should include the following information
Eliza Bechtold, Clara Burbano-Herrera, Zoë Grossi, Yves Haeck (eds.) *Echoes of Hate, Digital Communication, Populism and the Regulation of Hate Speech*, Human Rights in Context, Ghent University, 2025.

Human Rights in Context
solidarity with OTHERS
UNIVERSITY OF OXFORD
Bonavero Institute of Human Rights
idh Universitat de València Institut de Drets Humans
University of BRISTOL Human Rights Implementation Centre

Table of Contents

Editor's Note

Echoes of Hate: Digital Communication, Populism, and the Regulation of Hate Speech brings together a set of contributions that explore how digital technologies, political communication, and social tensions increasingly shape the presence of hate in contemporary public life. This book is the outcome of the international conference of the same title, held at Ghent University, which convened scholars, practitioners, and policy experts to examine the evolving challenges posed by digital communication in an era marked by polarisation and technological transformation. The chapters in this volume were written during a period of rapid change in the digital environment, when online platforms have become central to political debate, everyday interaction, and the spread of both information and disinformation.

A central message of this book is that online hate speech does not arise in isolation. It develops within a wider context marked by political polarization, economic anxiety, and the growing influence of populist movements. Digital platforms amplify these dynamics by rewarding messages that provoke strong emotional reactions. As several authors in this volume show, hateful or divisive messages often circulate faster and more widely than thoughtful or balanced communication. This makes digital spaces powerful arenas where identities are shaped, conflicts are intensified, and vulnerable groups can become targets of open hostility.

Another important theme in the book concerns the legal and regulatory challenges raised by online hate. Governments, especially in the European Union, have attempted to address these challenges through new regulations. These laws aim to increase transparency, strengthen content-moderation systems, and clarify the responsibilities of digital platforms. However, the authors also point out that laws alone cannot solve the problem. The border between harmful expression and protected speech remains difficult to define, and enforcement is complicated by the global nature of digital platforms. For this reason, the chapters stress the need for balanced approaches that protect users from harm while safeguarding freedom of expression.

Populism is another recurring focus in the volume. The contributors show how populist leaders use digital communication to construct a sharp divide between "the people" and "the enemies of the people." This rhetorical strategy often portrays minorities, immigrants, or political opponents as threats to the nation, thereby legitimizing negative stereotypes and encouraging online hostility. Digital tools—especially social media—allow such messages to be spread quickly and to reach large audiences without traditional editorial mediation. The result is an environment where emotional, confrontational, and sometimes hateful language becomes normalized.

A further key focus of the volume is gendered hate and online misogyny, which has become one of the most visible and troubling forms of digital hostility. Several chapters highlight how influential online figures, far-right women's groups, and anti-gender movements weaponize digital communication to attack women, undermine gender equality initiatives, and stigmatize sexual education. These contributions show that gender is not a side theme but a central axis of contemporary hate dynamics, deeply intertwined with broader political narratives and cultural anxieties.

Part I, "Governing Hate Online: Law, Responsibility, and Platform Power," examines the regulatory and institutional challenges raised by digital hostility. Lavinia Pedol analyzes the relationship between online hatred and offline violence, focusing on the EU's evolving legal response and the capacities—and limitations—of the Digital Services Act. Stephanie Reynolds' chapter investigates the unique political influence of platform owners, concentrating on Elon Musk's dual role as owner and user of X. Her contribution illustrates how private power and algorithmic visibility shape public debate, showing that online hate cannot be separated from the broader question of platform governance and responsibility.

Part II, "Populism and Digital Polarization," explores how divisive rhetoric and political communication fuel hostile online climates. Ibrahim Kurt discusses how populist strategies normalize hate by constructing sharp divisions between "the people" and "the other," and how social media magnifies these narratives. Francesca Cassano's chapter turns to the European Court of Human Rights, addressing the tension between protecting individuals from harmful expression and avoiding the chilling of legitimate speech. This part highlights how political messaging and legal doctrine together influence the boundaries of acceptable discourse.

Part III, "Gendered Hate and Online Misogyny," addresses one of the most visible and persistent forms of digital hostility. Elizabeth Pearson examines the rise of extreme misogynist communities, including the cultural influence of figures like Andrew Tate. Gwenaëlle Bauvois analyzes how far-right feminist groups strategically adopt the language of women's rights to advance exclusionary narratives. Katrien Jacobs and Katelijne Lievens study how anti-gender movements frame sexuality education as a threat, showing how fear-based rhetoric travels across digital spaces. These chapters reveal the gendered dimensions of online hate and the political projects it often serves.

Part IV, "Media, Journalism, and Countering Hate Narratives," turns toward information ecosystems and the role of journalists and fact-checkers. Carla Sentí Navarro discusses how disinformation targeting minorities corrodes democratic debate. Juliana da Cunha Mota provides a systematic review of fact-checking as a tool for addressing hate speech, assessing both its potential and its limits. Kezban

Karagöz offers an important perspective on journalism in exile, reflecting on the vulnerability of reporters who confront political hostility while navigating threats to personal safety.

Part V, "Emotions, Psychology, and the Human Experience of Hate," explores how people interpret and internalize hostile narratives. Aneta Szarfenberg documents how hate speech targeting individuals associated with the Gülen movement can move from digital labeling to real-world harms through administrative and legal decisions. Soraya Afzali discusses digital platforms as tools of influence in polarized political environments, emphasizing how emotions and identity shape political behavior. This final section underscores that hate speech is not only a legal or technological issue—it is also deeply psychological and experiential.

As editors, we hope this collection contributes to a deeper understanding of how digital communication and political developments influence the spread of hate speech. The authors offer valuable insights into both the risks and the opportunities created by today's digital environment. Their analyses remind us that protecting human dignity and democratic values requires constant attention and cooperation across disciplines.

We thank all contributors for their thoughtful work and the conference participants whose discussions helped shape this volume. We extend our sincere gratitude to Human Rights in Context, Ghent University; Bonavero Institute of Human Rights, University of Oxford; Human Rights Implementation Centre, University of Bristol, Institut de Drets Humans, Universitat de València; Solidarity with OTHERS. We also hope that this book encourages further research and public debate on how societies can respond effectively to hatred in the digital age, without losing sight of the essential freedoms on which open democracies depend.

Editors.

Part I

Governing Hate Online: Law, Responsibility, and Platform Power

From Online Hatred to Violence: The Architecture of Digital Hostility and the EU's Juridical Response

Lavinia Pedol, University of Milan

1. Introduction

In recent years, the issue of hate speech has gained increasing prominence within legal, political, and social debates, driven by the profound transformation of communicative spaces brought about by digital platforms (A. Siegel, 2020: 66 ss.). The speed at which online content is produced, shared, and disseminated has amplified not only opportunities for expression but also dynamics of hostility, discrimination, and verbal violence (Cavagnoli, 2022: 20 ss.). In this context, identifying a legally sound and practically workable notion of hate speech represents a significant challenge, as it involves regulatory, technological, and socio-cultural dimensions that are deeply intertwined.

The growing prevalence of online hate speech has prompted international institutions – and the European Union in particular – to reflect on the role of major digital intermediaries, which are called upon to balance freedom of expression, the protection of fundamental rights, and responsibility in content moderation (Ferrucci, 2019; Goisis, 2019; D'Amico, 2023). EU initiatives, ranging from the non-binding Code of Conduct on Countering Illegal Hate Speech Online to the recent adoption of the Digital Services Act (DSA), reflect an attempt to develop a regulatory framework capable of responding to the transnational and technically complex nature of digital platforms (Wilman, Kaleda, 2024; Correra, 2024).
However, significant concerns remain between the regulatory ambition of the DSA and the operational reality of digital ecosystems. While the new Regulation introduces innovative tools for addressing illegal content and ensuring algorithmic transparency, doubts persist regarding the actual capacity of the obligations imposed to align with the economic and technical dynamics of the private actors that dominate the market.

This contribution, therefore, aims to provide a critical analysis of the phenomenon of online hate speech, reconstruct the framework of the European response, and assess the actual scope of the DSA in bridging the gap between normative objectives and concrete implementation. In this context, particular attention is also devoted to the role of criminal law, both as a traditional instrument for addressing the most serious forms of hate speech and as a benchmark for evaluating the proportionality and effectiveness of regulatory measures in the digital environment. The objective is to offer a systematic examination that brings into dialogue the theoretical, legal, and regulatory dimensions of the topic, thereby contributing to the broader debate on the governance of digital content in an ever-evolving environment.

2. Hate Speech: An Attempt to Define a Complex Phenomenon

Few concepts in contemporary public discourse are as elusive and disputed as hate speech (see, among many others, Assimakopouloss, 2020, Anderson & Barnes, 2022; Ermida, 2023). Its definition continues to resist consensus, reflecting the social, moral, and theoretical tensions that underlie its many forms (Weinman, 2006). The difficulty stems from the fluid boundaries between intolerance, verbal aggression, and outright incitement to discrimination (Ermida, 2023: 35 ss.). Some scholars equate hate speech with offensive or abusive language; others restrict it to expressions that explicitly promote hostility (Neller, 2023: 28 ss.; Duff & Marshall, 2028: 115). Each perspective captures only part of a broader, far more intricate reality. The resulting semantic uncertainty has long hindered the construction of a coherent and stable theoretical foundation.

At the heart of this debate lies a delicate equilibrium between two competing imperatives: the safeguarding of freedom of expression – an indispensable cornerstone of democratic life – and the protection of equality and human dignity, which hate speech directly undermines (Crusschina & Gianollo, 2024). The thin line separating legitimate opinion from communicative harm remains the subject of intense philosophical and legal reflection.

The diversity of national legal systems compounds the problem, each guided by its own historical and cultural sensibilities. What is protected speech in one jurisdiction may constitute a criminal offence in another. The values at stake – freedom, equality, and dignity – acquire different meanings and relative weights depending on the sociopolitical context. For this reason, any attempt to elaborate a single, exhaustive definition of hate speech is inevitably fraught with difficulty (Cabo Isasi and García Juanatey 2016).

The modern category of hate speech took shape in post-war jurisprudence and philosophical debate (Ruscher, 2024: 6 ss.). It gained visibility in the United States during the 1970s, when university campuses became sites of confrontation over racial and ideological discrimination (Di Lisio, Sorrentino, Trezza, 2022: 50 ss; Guillèn-Nieto, 2023: 109 ss.). The landmark case Brandenburg v. Ohio (1969) crystallised the fundamental tension between advocacy and incitement (Galluccio, 2020: 286; Guillèn-Nieto, 2023: 40-42). The American model, grounded in the First Amendment's broad protection of expression, diverges sharply from the European approach, which treats freedom of speech as a right to be balanced against other fundamental rights (Galluccio, 2020: 196; Guillèn-Nieto, 2023: 23 ss.).

In Europe, the recognition of the Internet as a global conduit for hatred and incitement to violence has prompted the gradual evolution of a legal and regulatory framework designed to protect individuals' fundamental rights (Ruscher, 2024: 6 ss.). These efforts build upon the post-war human-rights architecture. Article 7 of the Universal Declaration of Human Rights (1948) guarantees protection against discrimination and incitement, while Article 20 of the International Covenant on Civil and Political Rights (1966) explicitly prohibits advocacy of national, racial, or religious hatred that constitutes incitement to hostility or violence.

Within the Council of Europe, the European Commission against Racism and Intolerance (ECRI) has played a central role in shaping this normative response. Its recommendations – beginning with General Policy Recommendation No. 1(1996) and culminating in No. 15 (2015) – chart a widening of concern from racism and xenophobia to gender, disability, and sexual orientation. The 2015 document provides one of the most comprehensive definitions to date, describing hate speech as any advocacy, promotion, or justification of hatred, denigration, or defamation based on identity characteristics such as race, colour, language, religion, nationality, or sex (Bello & Scudieri, 2022: 8 ss.).

Within the framework of the Council of Europe, the European Commission against Racism and Intolerance (ECRI) has played a crucial role through a series of recommendations promoting equality and combating discrimination. These include General Policy Recommendation No. 1 (1996), on combating racism, xenophobia, antisemitism, and intolerance; No. 2 (1997), condemning all forms of discrimination based on ethnicity, language, religion, nationality, sex, sexual orientation, or gender identity; and No. 7 (2002), emphasizing the urgency of legislative measures to counter racial discrimination.

On 30 October 1997, the ECRI also adopted a specific recommendation against hate speech, further developed in General Policy Recommendation No. 15 (2015), which defines hate speech as any advocacy, promotion, or incitement of denigration, hatred, or defamation of individuals or groups, as well as the justification of such expressions, based on characteristics such as race, colour, language, religion, nationality, ethnic origin, age, disability, sex, gender identity, sexual orientation, or other personal status.

At the international level, the Additional Protocol to the Convention on Cybercrime (Budapest Convention) – in force since 2006 – is also significant, as it criminalizes specific racist or xenophobic acts committed via computer networks. Equally fundamental are the provisions of the European Convention on Human Rights (ECHR), particularly Article 10, which protects freedom of expression while balancing it against the need to prevent hate speech and incitement to violence.

The normative framework is further enriched by the European Social Charter, Article 19 of the Universal Declaration of Human Rights, Article 13(5) of the American Convention on Human Rights, as well as numerous declarations and recommendations of the Council of Europe and the United Nations – all converging in condemning hate propaganda and promoting responsible freedom of expression.

In general terms, the concept of hate speech encompasses all forms of expression manifesting hostility or intolerance toward individuals or social groups, with the potential risk of generating or amplifying violent reactions (Pino, 2008). Hate speech may take different forms: verbally, it manifests using derogatory labels for groups, dehumanizing metaphors, or negative exemplifications; visually, it may appear through threatening symbols or imagery (Culpeper, 2021; Ruscher, 2024). Hatred may be expressed in face-to-face contexts or through mass media, including websites and social media platforms, characterized by their broad public accessibility.

UNESCO identifies several features that distinguish online hatred from its traditional forms: its endurance across time and platforms; its replication capacity; the anonymity that digital tools afford; and its transnational character, which often renders national legal instruments ineffective (Ziccardi & Perri, 2022). These characteristics make online hate not only more pervasive but more insidious, embedding intolerance within the very architecture of communication.

The consequences are profound. For individuals, exposure to sustained hostility generates anxiety, fear, and alienation, while reinforcing negative stereotypes and self-perceptions (Neller, 2023: 29). For society, the normalisation of hate corrodes trust and civic dialogue, polarises communities, and fosters environments conducive to harassment and, ultimately, to hate crimes (Goisis, 2019). Lawmakers face the daunting challenge of distinguishing genuine hate advocacy from satire or legitimate opinion, of defining the responsibilities of digital platforms, and of redrawing the boundaries between freedom and harm in a world where speech is instantaneous and borderless.

3. The Role of Online Hate within Digital Platforms

Once relegated to the margins of the Internet, online hatred has become a structural component of the digital ecosystem (A. Siegel, 2020: 56 ss.). No longer confined to extremist forums, it now permeates mainstream social networks and comment sections, shaping the tone of public debate and the forms of political engagement (Irene Spigno, 2018: 19). The Internet thus emerges as a paradoxical arena where freedom of expression and human dignity coexist in constant tension, posing a challenge to the democratic resilience of contemporary societies (Francisco Javier Ansuategui Roig, 2017: 35 ss; Conti, 2018).

Hatred can be understood as a complex psychological and social phenomenon composed of three elements: rejection of intimacy, aversion towards what is perceived as different, passion – emotions such as anger or fear directed at what is perceived as threatening, and commitment, the will to devalue or destroy the other (Spigno, 2018: 19 ss.). Legally and morally, hatred differs from mere dislike: it is a deep, prejudice-driven hostility capable of fuelling resentment, discrimination, and violence (Sternberg, 2005: 45-58). Once disseminated, it can evolve into hate crimes motivated by bias against the victim’s identity.

The causes of hatred are manifold: personal prejudice, social tensions, symbolic conflicts between groups, and structural inequalities (Cortese, 2006). In its contemporary form, hatred is often normalised through everyday language, rhetorical dramatization of evil, and processes of blame and stigmatization (Waldron, 2014). No longer limited to overt acts of violence, it thrives through online discursive practices, comments, memes, and posts that are produced and circulated daily by ordinary users (Frosini, 2024).

Digital technologies have turned the Internet into a global public sphere in which verbal aggression has become routine (Boccia Artieri & Marinelli, 2018: 350-360; Ziccardi & Perri, 2022: 95 ss.). Social media, initially conceived as tools of connection and democratic participation, now function as powerful amplifiers of hostility (Parekh, 2012: 40-56). Online platforms admittedly keep the world open to millions and allow communicators to creatively interact on a global scale, but they also set traps for the unwary user (Ermida, 2023: 4; Frosini, 2024). Algorithms designed to maximise engagement favour emotionally charged content, nurturing polarisation and radicalisation (Blanks, 2010; Duffy, Poell & Nieborg, 2019; Ermida, 2023; Tommasi, 2023). Within such closed communicative loops, commonly described as filter bubbles and echo chambers (Corazza, 2022), users encounter mainly opinions that reinforce their own, thereby intensifying intolerance (Correra, 2024: 177).

The Internet's accessibility has transformed every user into both consumer and producer of content (Benkler, 2006; Carr & Hayes, 2015: 8; Conti, 2018). This unprecedented participatory model has broadened the public sphere but also diluted the sense of responsibility attached to speech. The illusion of anonymity fosters disinhibition; empathy fades behind the screen. What is written online persists, often indefinitely, leaving tangible traces of hostility and fear (Benesh, 2014).

Consequently, the language of the web has become saturated with derision and aggression. Insults, body shaming, and disparaging comments are no longer exceptions but integral features of digital communication (Buyse, 2014: 778-796). These expressions both reflect and reinforce existing social hierarchies, disproportionately affecting vulnerable groups and minorities. Perpetrators, meanwhile, derive from such behaviour a false sense of superiority and belonging (Neller, 2023).

Contemporary hatred manifests in two main trajectories: the traditional, directed against minority groups perceived as "different," and the interpersonal, aimed at specific individuals (Cruschina & Gianollo, 2024: 3 ss.). This discussion focuses on the former – online hostility towards social minorities in conditions of particular vulnerability. These so-called target groups include communities historically exposed to oppression or discrimination: women, people with disabilities, members of the LGBTQIA+ community, and ethnic or religious minorities (Bell, 2002: 181; Brown, 2017: 23-25; Bianchi, 2024). Bias indicators – analytical tools used to detect discriminatory intent – reveal how online campaigns of denigration often serve not only to humiliate victims but also to reaffirm the aggressor's sense of power and belonging (Goisis, 2019).

Online hatred thus performs a social function: it draws symbolic boundaries between "us" and "them," legitimising exclusion and reinforcing collective identities built on opposition (A. Siegel, 2023).

The immediacy of digital communication and the illusion of impunity intensify this phenomenon. The Internet is perceived as a realm of expanded freedom, where one can express anger or prejudice without consequence. Yet this perceived liberty frequently translates into new forms of verbal violence. Incivility, once marginal, now shapes the tone of online discourse, often under the guise of humour or authenticity (Ruscher, 2024).

Although online hatred is not new, its recent escalation is striking. Global events such as migration crises, terrorist attacks, political polarisation, and the COVID-19 pandemic have amplified speculative narratives grounded in fear and misinformation (Di Lisio, Sorrentino, Trezza, 2022; Retta, 2024: 233-280; Safina: 2024: 111-149). The overabundance of news, combined with the absence of editorial mediation, has turned social media into fertile ground for hostility. Increasingly, hate speech – whether explicit or subtle – has become embedded in public discourse itself, blurring the line between political opinion and aggression (Waldron, 2021).

As Rega and Marchetti (2019) observe, forms of communicative incivility are no longer marginal; they are amplified and normalised through participatory digital dynamics. Yet hate speech occupies a paradoxical position: it can represent both an exercise of free expression and a violation of human dignity. Because it touches the constitutional foundations of democracy, it demands a legal and ethical response (Pollicino & De Gregorio, 2019).

The digital transformation has thus redefined the dynamics of hostility, endowing hate with unprecedented reach and complexity. Confronting it requires more than censorship or criminalisation: it calls for critical reflection on the cultural, educational, and institutional tools necessary to prevent its diffusion. In this light, the struggle against online hatred emerges as one of the most urgent and challenging tasks facing democratic societies today.

4. The European Union's Response to Online Hate Speech

In light of the continuous evolution of hateful content on the Internet and the parallel transformation of social contexts, it is necessary to reflect on which strategies are most effective in containing and countering the phenomenon of online hate (Scamuzzi, Belluati, Caielli, Cepernich, Patti, Stecca & Tipaldo, 2021).

Criminal law (and liability law more generally) is confronted with complex questions: to what extent can a digital platform be held liable for unlawful content (including content amounting to hate speech)? What "notice-and-action" obligations does the platform bear? What role is played by user reporting?

In an attempt to provide an organic, multi-level response to the dissemination of unlawful online content, the European Union has progressively developed a normative and para-normative framework aimed at reinforcing the accountability of online platforms and, more broadly, of digital communication intermediaries (Bassini, 2021). The objective has been to create a synergistic alliance between private actors and European institutions (Iannotti Della Valle, 2020: 142 ss).

A first fundamental reference point is Directive 2000/31/EC on electronic commerce (the e-Commerce Directive), transposed in Italy by Legislative Decree No. 70 of 9 April 2003, which laid the foundation for online commerce and established rules governing intermediary liability. The directive introduced the safe-harbour principle – namely, a general exemption from liability for Internet Service Providers (ISPs) concerning user-generated content (Correra, 2024). This principle rests on the assumption that intermediaries are not subject to a duty of prior control over transmitted or stored information and, consequently, prohibits Member States from imposing general monitoring obligations. The duty to intervene arises only once actual knowledge of the unlawfulness of the content has been obtained ("notice and takedown").

Parallel developments occurred at the international level. The Additional Protocol to the Budapest Convention on Cybercrime (2001) broadened the scope of cooperation among States in combating cybercrime, laying the groundwork for increased attention to online hate content and incitement to violence.

From the second half of the 2010s onward, the European Commission promoted several soft-law initiatives and co-regulated self-regulation mechanisms intended to enhance the responsibility of digital platforms. In this context, the "Code of Conduct on Countering Illegal Hate Speech Online," signed in 2016 between the Commission and major global technology companies (Google, Facebook, Twitter and Microsoft), is of particular importance (Di Lisio, Sorrentino, Trezza, 2022; Correra, 2024). Based on voluntary commitments, the Code established structured cooperation between public institutions and private operators to ensure faster and more effective review and removal of content reported as hate speech.

The experience of the Code of Conduct fits within a broader European strategy of multi-level governance, which also includes the "EU Internet Forum" project (2015), designed to coordinate national governments, Europol and technology companies in combating terrorism and hate-related online content. This approach reflects the Union's intention to construct a public–private cooperation network capable of serving as a regulatory umbrella for national policies (Scamuzzi et al., 2021).

Despite the progress made in terms of raising awareness and increasing content-moderation procedures, soft-law instruments have proven fragile, being rooted in voluntary and non-binding commitments. Further challenges stem from the linguistic and cultural diversity of the European context – which complicates the development of a uniform definition of "hate content" – and from the persistent reluctance of certain Member States to implement EU recommendations consistently, at times perceived as intrusive vis-à-vis national sovereignty.

To overcome these limitations, the Commission has gradually oriented its action toward the adoption of hard-law instruments capable of converting the good practices developed under the Codes of Conduct into legally binding obligations. In this respect, a significant step was taken with the revision of the Audiovisual Media Services Directive (Directive 2018/1808/EU), which extended monitoring and removal obligations to video-sharing platforms, imposing specific measures to counter the dissemination of illegal content, particularly content inciting hatred or violence.

These regulatory developments find further consolidation in the Digital Services Act (Regulation (EU) 2022/2065), which marks the transition from a logic of mere self-regulation to a model of shared and transparent responsibility (Caggiano, 2021: 4 ss.). The DSA introduces obligations of due diligence, algorithmic transparency and notice-and-action systems aimed at safeguarding fundamental rights within the European digital sphere (Bolognini, Pelino, Scialdone, 2023).

The Digital Services Act (DSA) maintains systematic continuity with Directive 2000/31/EC, of which it represents a substantial evolution. It establishes a more articulated and stringent regulatory framework for providers of information-hosting services – including large online platforms – redefining their duties and responsibilities in terms of transparency, traceability, and systemic risk management (Tommasi, 2023).

The reform leading to the adoption of the DSA forms part of the Union's broader project to build a safe, predictable and trusted online environment, within which the protection of fundamental rights, legal certainty and fair competition coexist in a dynamic equilibrium between technological innovation and legal oversight (Pollicino & De Gregorio, 2019; Correra, 2024).

The DSA contains provisions specifically targeting unlawful online content, including online hate speech, through the imposition of new obligations on digital service providers. While maintaining the core logic of the 2000 e-Commerce Directive – namely, the exemption of platforms from liability for user-generated unlawful acts and the prohibition of ex ante content monitoring – the DSA combines these principles with a series of procedural and substantive obligations. These include the establishment of internal mechanisms for the removal of unlawful content and the requirement to provide reasons for, and allow appeals against, decisions taken.

In principle, the provider remains exempt from liability until it gains actual knowledge of the unlawful content: this moment marks the dies a quo from which a duty of prompt action arises to identify the source, block dissemination, and prevent access. This regime of so-called "conditional liability" is implemented through various measures, including quicker and more structured procedures for eliminating illegal items or products, user reporting mechanisms, and the introduction of specialised service categories required to inform the judicial authority of the Member State where the service is established whenever a suspected offence poses a threat to the life or safety of one or more persons (Article 18).

To ensure the effective suppression of unlawful online content, the DSA adopts a decentralised approach involving cooperation between public authorities and organisations or stakeholders with recognised expertise. Within this framework, content oversight is entrusted to entities defined as "trusted flaggers" (Article 20): authorities or NGOs with proven experience and competence in detecting and reporting unlawful online content. Reports submitted by such actors must receive priority handling to guarantee timely and effective platform intervention (De Gregorio, 2021).

Article 16 introduces notice-and-action tools, requiring all hosting services – not only platforms – to implement simple and accessible mechanisms enabling both users and authorities to report problematic content, including hate speech. Platforms must respond promptly, assessing reports accurately and transparently and informing the reporting party of the decision taken.

Under Article 15, when a platform decides to remove unlawful content or restrict access to it, it must provide the user with a clear, precise and comprehensible explanation. Furthermore, platforms must establish internal complaint mechanisms allowing users to contest removals they deem unjustified, thus ensuring a balance between freedom of expression and online safety (Article 17).

Particularly relevant are Articles 34 and 35, which focus on Very Large Online Platforms (VLOPs) – platforms with more than 45 million users in the European Union. These platforms are required to identify and mitigate systemic risks arising from their services, with special attention to content-moderation processes and algorithmic transparency. The aim is to enhance platform accountability, ensuring that the scale of their influence does not compromise user safety or fundamental rights.

VLOPs – defined through a quantitative criterion – must periodically conduct assessments of systemic risks associated with their services, evaluating them according to the gravity and probability of potential harms. Notably, in cases of non-compliance, Member States may impose financial penalties on large platforms that may reach up to 6% of their annual turnover (Article 52).

Ensuring the prompt and rigorous interruption of the flow of unlawful content thus constitutes one of the central duties imposed on digital platforms, reflecting the growing expectation that intermediaries actively contribute to safeguarding the integrity of the online information ecosystem.

5. The DSA Between Ambition and Reality

The Digital Services Act (DSA) constitutes one of the European Union's most ambitious remarks in the regulation of digital platforms, representing a comprehensive attempt to redesign the architecture of digital governance (Tommasi, 2023). Nonetheless, its actual capacity to shape the European information and communication ecosystem remains subject to empirical and doctrinal assessment.

Given the recency of its application, it is still premature to formulate definitive judgments concerning its systemic effects; yet some initial interpretative and operational challenges can already be identified.

As previously noted, the DSA introduces innovative mechanisms of transparency and accountability for providers of intermediary services, with the stated aim of ensuring a safer, fairer, and more rights-respecting digital environment. Despite its ambitious regulatory design, the framework presents several areas of uncertainty that risk undermining the uniformity of its implementation.

A first critical issue concerns Article 3 of the DSA, where the definition of unlawful content appears intentionally broad and indeterminate. Although such openness serves the purpose of allowing interpretative flexibility, it risks resulting in heterogeneous applications across Member States and in conferring excessive discretionary power upon platforms in content removal. It therefore seems desirable to adopt EU-wide interpretative guidelines capable of providing clearer criteria for identifying sanctionable conduct.

A second problematic aspect relates to the identification of the authors of online hate speech. In practice, enforcement is hampered by the fact that access to users' identifying data remains contingent on the cooperation of platforms, which only rarely provide the requested information, notwithstanding the relevant provisions of the GDPR on the processing of personal data. The widespread use of pseudonyms, fake accounts, and anonymisation tools (VPNs, the Tor network, proxy servers) further complicates traceability, making it necessary to adopt measures aimed at preventing the abuse of digital anonymity (Woods & Ruscher, 2021; Frosini, 2024: 120 ss.).

A third interpretative challenge concerns the balancing of freedom of expression with the public interest in collective security. The progressive privatisation of content control – through the delegation to platforms of moderation powers – risks resulting in a form of discretionary enforcement which, in the absence of effective judicial oversight, may compromise the protection of fundamental rights. Such dynamics can generate a chilling effect, prompting users to self-censor for fear of sanctions or arbitrary removal.

The boundary between hate speech and hate action has become increasingly blurred, thus justifying restrictive intervention by platforms in relation to content that incites hatred or violence. Although freedom of expression is a cornerstone of democratic systems, it cannot be exercised without limits, particularly where hate speech produces tangible social effects and contributes to the proliferation of violence, discrimination and insecurity. The events in Halle (2019) and Christchurch (2019) represent emblematic cases in which online radicalisation manifested in real-world violence, demonstrating that digital hatred transcends the confines of cyberspace.

This underscores the need to assign platforms a proactive role in countering unlawful content, especially where it constitutes incitement to hatred or violence. The challenge for the DSA – conceived as a new digital constitution for the EU

– is to reconcile the protection of collective security with the safeguarding of pluralism and open democratic debate.

From this perspective, it is essential that platforms continue to invest in artificial intelligence technologies for the automatic detection of hateful content, improving algorithmic accuracy so as to avoid over-blocking and to ensure proportionate and non-discriminatory ex post review. However, technological solutions alone are insufficient: it is necessary to promote a widespread culture of digital responsibility, grounded in critical awareness of algorithmic functioning, content-amplification mechanisms and the social consequences of online interactions.

Only through the integration of legal, technical and educational instruments will it be possible to achieve the DSA's ultimate objective: a European digital ecosystem that is pluralistic, secure and respectful of human dignity.

REFERENCES

Ansuategui Roig F.J., *Libertà di espressione, discorsi d'odio, soggetti vulnerabili: paradigmi e nuove frontiere*, Ars interpretandi (ISSN 1722-8352) Fascicolo 1, gennaio-giugno 2017, il Mulino.

Assimakopouloss L., *Incitement to discriminatory hatred, illocution and perlocution*, Pragmatics and society, 11(2), 2020.

Balkin J. M., *The Future of Free Expression*, in a Digital Age, 36/2009.

Bassini M., Libertà di espressione e social network, tra nuovi "spazi pubblici" e "poteri privati". Spunti di comparazione, in *Rivista italiana di informatica e diritto*, 2/2021.

Bell J., *Policing Hatred: Law Enforcement, Civil Rights, and Hate Crime*, New York and London: New York University Press, 2002.

Bello B. G., Scudieri L., *Discorsi d'odio online. Spunti per un dibattito interdisciplinare*, in Bello B. G., Scudieri L., (a cura di) L'odio online: forme, prevenzione e contrasto, Giappichelli, Torino, 2022.

Benkler Y., *The Wealth of Networks: How social production transforms markets and freedom*, Yale University Press, 2006.

Bianchi C., *Hate speech il lato oscuro del linguaggio*, Laterza, 2021.

Bolognini L., Pelino E., Scialdone M., *Digital Service Act e digital market act,* Giuffrè, Milano, 2023.

Brown A., The who? Question in the Hate speech debate: part 2: functional and democratic approaches, *Canadian journal of law and jurisprudence* 30, 1/2017.

Buyse A., Words of Violence: “fear speech” or how violent conflict escalation relates to the freedom of expression, *Human Rights Quarterly*, 4/2014.

Cavagnoli S., Le parole fanno male. E anche le immagini, in Bello B. G., Scudieri L., (a cura di) *L’odio online: forme, prevenzione e contrasto*, Giappichelli, Torino, 2022.

Conti G.L., *Manifestazione del Pensiero attraverso la rete e trasformazione della libertà di espressione: c’è ancora da ballare per strada?*, in Riv. Aic, 4/2018.

Corazza P., *Filter bubbles e echo chambers: pre-digital origins and elements of novelty. Reflections from a media education perspective, formazione & insegnamento*, 2022.

Correra A., *La strategia dell’Unione europea controi i discorsi d’odio online: il ruolo delle piattaforme digital e la ricercar di un equilibrio tra la tutela della libertà di espressione e un sistema di enforcement in tempi e contesti di crisi,* in Emma A. Imparato e G. Giorgini Pignatiello (a cura di) La libertà di espressione nel diritto comparato tra stato di diritto e stati di emergenza, Giappichelli, Torino, 2024.

Cruschina S., Gianollo C., *An investigation of hate speech in Italian. Use, identification, and perception*, Helsinki University Press, 2024.

Culpeper J., *Impoliteness: Using language to cause offence*, Cambridge: University press, 2021.

D’Amico M., *Parole che separano, linguaggio, costituzione, diritti*, Cortina, Milano, 2023.

De Gregorio G., *The Digital Services Act: A paradigmatic example of European digital constitutionalism*, in Diritti comparati, 2017.

Di Lisio M., Sorrentino R., Trezza D., *Platformization hate. Patterns and algorithmic bias of verbal violence on social media*, Mediascapes journal 20/2022.

Duff R., Marshall S.E., Criminalizing Hate?, in *Hate, Politics, Law: Critical Perspectives on Combating Hate*, (ed.) Brudholm T., Johansen B.S., New York: Oxford University Press, 2018.

Ermida I., *Hate speech in social media*, Palgrave Macmillan, 2023.

Ferrucci F., *L’hate speech, l’odio nel discorso pubblico*, Robin, Roma, 2019.

Frosini T.E., La libertà di manifestazione del pensiero nell’era di Internet, in Emma A. Imparato e G. Giorgini Pignatiello (a cura di) *La libertà di espressione nel diritto comparato tra stato di diritto e stati di emergenza*, Giappichelli, Torino, 2024.

Galluccio A., *Punire la parola pericolosa*, Giuffrè, Milano, 2020.

Goisis L., Crimini d’odio. *Discriminazioni e giustizia penale*, Jovene editore, 2019.

Guillèn-Nieto V., ‘Hate speech. Linguistic perspectives’, *Foundations in language and law*, Vol. 2, 2023.

Herz M., Molnar P., *The content and the context of hate speech. Rethinking regulations and responses*, Cambridge University Press, 2012.

Neller J., *Stirring up hatred. Myth, identity and order in the regulation of hate speech*, Palgrave Macmillan, 2023.

Parekh B., Is there a Case for banning hate speech?, in Herz M., Molnar P., (ed.) *The content and context of hate speech. Rethinking regulation and responses*, New York: Cambridge University Press, 2012.

Pollicino O., De Gregorio G., Hate speech: una prospettiva di diritto costituzionale comparato, in *Giornale di diritto amministrativo*, 4/2019.

Ruscher J.B., *Hate speech*, Cambridge University Press, 2024.

Siegel A.A., *Online hate speech*, Cambridge University Press, 2023.

Spigno I., *Discorsi d’odio. Modelli costituzionali a confronto*, Giuffrè editore, Milano, 2018.

Spigno I., Quando la parola discrimina: l’ingresso dell’art. 14 Cedu nella giurisprudenza europea sui discorsi d’odio, in *Quaderni costituzionali* (ISSN 0392-6664) Fascicolo 2, giugno 2021, il Mulino;

Stenberg R.J., (a cura di) The psychology of hate, Washington d.c., *American Psychological Association*, 2005.

Tommasi S., *The risk of discrimination in the digital market. From the digital services act to the future*, Springer, 2023.

Vasino G., *Censura “privata” e contrasto all’hate speech nell’era delle Internet Platforms*, federalismi.it, n.4/2023.

Waldron J., *The harm in hate speech*, Cambridge: Harvard University Press, 2014.

Walker S., Hate Speech. *The History of an American Controversy*, Lincoln, 1994.

Weinman M., State Speech vs. Hate Speech: What to do about words that wound?, *Essays in Philosophy* 7, 1/2006.

Weinstein J., *Extreme Speech and Democracy*; Oxford: Oxford University Press, 2009.

Ziccardi G., Perri P., L’odio online tra profilazione, big data e protezione dei dati personali, in Bello B. G., Scudieri L., (a cura di) *L’odio online: forme, prevenzione e contrasto*, Giappichelli, Torino, 2022.

Profiting from Hate or Participating in Debate? Critiquing the Unique Political Reach of Elon Musk as Owner-User of X

Stephanie Reynolds, University of Liverpool

Introduction

Over the last few years, Elon Musk has used his X account to make what many would view as increasingly populist interventions in European politics. His posts have, for instance, been accused of "fan[ning] the flames of unrest and rioting across the UK" (Griffin, 2024). Interestingly, other than a brief stint as a public actor whilst heading up the constitutionally questionable (Raul, 2025; Sneed, 2025; Moynihan, 2025) US Department of Government Efficiency (DOGE), Musk's political comments are usually presented as those of a concerned private citizen (Musk, n.d.-a). Yet as the world's richest person (Forbes, n.d.), and owner – rather than mere user – of X, Musk is no ordinary participant in online debate. Indeed, while a central concern about social media platforms is their algorithmic fostering of echo chambers (Willis, 2020), Musk's position as X owner, with 220 million followers, means such filtering is less likely to affect his own posts. In short, Musk's offline power extends his online reach, furnishing him with an extensive platform to voice his often-controversial views and to boost those of others.

Accordingly, Musk's unique position as owner-user of X offers a useful lens through which to explore the overarching themes of the Echoes of Hate conference. Section one of this paper thus outlines some of Musk's choices as X owner, his contentious interventions as X user, and some of the allegations levelled at him about the link between these online activities and offline violence. It posits that his privileged capacity to amplify populist rhetoric places him in a position of responsibility at the nexus between digital hate speech and hate crime, regardless of whether he commits such acts himself. Section two argues that Musk's ability to amplify populist messaging is not only underpinned by his online interjections. The reactions to Musk's online speech by established constitutional actors, as a result of his societal power, cement rather than counteract the megaphone Musk can offer his chosen narratives. Musk's consequent ability to operate as a populist leader (Musk, n.d.-b) throws into sharp focus broader concerns about unchecked private power, which operates outwith the checks and balances of public constitutional frameworks whilst still drawing on established tenets of liberal constitutionalism to legitimise and enable its dominance. These findings inform section three's assessments of the UK's regulatory response to online hate speech. Examination of whether Musk's posts leave him liable to conviction for any criminal offences, as a user, under the UK's Online Safety Act 2023, or impose any responsibility on him as an owner demonstrate the weaknesses in the UK legislation. Yet, it also highlights the hamstrung nature of any regulatory attempt to balance curbs on populist narratives against the need to protect free speech. Consequently, the section asks whether constitutional and societal attempts to address the specific drivers of Musk's – and others' – private societal power might offer a surer means of undercutting the seductive nature of the narratives he arguably perpetuates, and which also thrive among ordinary social media users. In other words, addressing societal inequality and practical disenfranchisement might be more effective than futile regulatory attempts to square the free speech circle.

1. A position of responsibility at the nexus of online hate speech and hate crime? An overview of Musk's contentious interventions in UK politics

When Musk became the owner of X – then Twitter – in October 2022, he declared that the "bird [was] freed," later underscoring that the platform's central purpose was to be a "global town square – from the people, for the people" (Musk, n.d.-c). For Musk, this meant re-instating accounts of certain commentators who had been banned from the platform under its content moderation policies, since, he argued, X should house all political views (Rawlinson, 2018; Anon, 2025; Statista, n.d.-a). Alongside these changes, Musk disbanded the platform's Trust and Safety Council (Anon, 2022)– an advisory group comprising nearly 100 independent civil, human rights, and other organisations – and fired Twitter's Chief Executive in charge of Trust and Safety (Ortutay et al., 2022; Brewster, 2024).

Although it is methodologically difficult to track the relationship between social media activity and offline incidents, it is at least arguable that some of these decisions were contributing factors to recent political unrest in the UK (Spring, 2024a). This includes the rioting that took place across the country following the tragic killing of children in Southport during a Taylor Swift-themed dance class (Merseyside Police, 2024). In the aftermath of the attack, misinformation quickly spread on social media that it had been carried out by a Muslim refugee who had arrived by small boat in 2023 (Spring, 2024b). In fact, the perpetrator, Axel Rudakubana, was born in Cardiff and had been living in the Southport area since 2013. Nevertheless, re-instated X user, and "self-proclaimed misogynist" (BBC News, 2025a) Andrew Tate, who at the time of the attacks had around 9 million followers (BBC News, 2025a), was reported as posting that the attacker was an "illegal immigrant" and that people needed to "wake up" (Cumming, 2024; Milmo & Quinn, 2024; Browning, 2024). On the day the 2024 riots began, former English Defence League leader and re-instated X user, Stephen Yaxley-Lennon – known as "Tommy Robinson" (Lindsay, 2018) – used the platform to push an anti-Muslim narrative and to argue that "[p]eople need to rise up…our daughters are being butchered" (Robinson, n.d.-a). The post received over 630,000 views.

This pales into insignificance when compared with the total number of interactions with Robinson's extensive X postings over the course of the riots. A video, posted the day they began, arguing that there was "evidence" that "[I]slam is a mental illness" garnered 1.5 million views and 41,000 "likes," while a video he posted on 2nd August, when the riots were ongoing, of a police station being destroyed as "the people revolt" was watched 5.4 million times and "liked" 38,000 times (Community Note on X, n.d.; Robinson, n.d.-b). On the same post, Robinson tagged Prime Minister Keir Starmer, telling him that he "should have listened." The next day, Robinson used X to "report" that two "#EnoughisEnough" protestors had been stabbed by Muslims (Robinson, n.d.-c). This post was viewed 2.7 million times. Fact-checking service FullFact.org subsequently reported that this information was false (Full Fact, n.d.). While it cannot be suggested that Robinson single-handedly instigated the rioting, he clearly did not shy away from supporting those engaged in them. Indeed, he explicitly accused Starmer of calling protestors "thugs" (Spring, 2024b), when they were "justified in their anger" (Milmo & Quinn, 2024; Unknown, 2024). Certainly, Robinson became something of a figurehead during the riots, with protestors heard chanting his name during rallies (Marshall, 2024).

Crucially, various media outlets (Griffin, 2024; The Times, 2024) and NGOs have argued that, as well as reinstating Robinson's account, changes made by Musk to X when he became owner resulted in the greater visibility of Robinson's content (Centre for Countering Digital Hate, 2024). Robinson holds a "blue tick" on the platform, which was identified even prior to Musk's takeover as lending credibility to account content regardless of its actual veracity (House of Commons Digital, Culture, Media and Sport Committee, 2020). Musk, however, explicitly linked the "blue tick" to X's "premium service" and the consequent prioritisation of content in algorithmic rankings (X, n.d.-a). Robinson himself clearly considers this service to be beneficial, frequently posting statistics on the interactions his content receives on the platform (TrobinsonNewEra, n.d.-d; Centre for Countering Digital Hate, 2024). Even without this paid-for benefit, X's algorithmic coding might favour content like Robinson's. An analysis of X's open-source code, conducted by Amnesty International, suggested that "the platform's content ranking algorithms prioritise the type of content that can spread misinformation and hate, with deeply inadequate safeguards to prevent human rights abuses…these design choices significantly exacerbated human rights risks for racialised communities in the wake of the Southport riots"

Similarly, the Centre for Countering Digital Hate has argued that “X’s algorithm is designed to reward controversial content that gets views and engagement” (Centre for Countering Digital Hate, 2024). The result, they posited, was that Robinson’s posts, which linked the Southport murders to Muslims, were viewed 13 million times. Crucially, as the CDH highlights, engagement with Robinson’s posts increased dramatically during the riots. While he accrued an average of 11.1 million daily views prior to the Southport attacks, he garnered 54.3 million during the unrest (Centre for Countering Digital Hate, 2024). This suggests a feedback loop between online content and real-world activity. While it might be argued that Robinson’s posts were a driver of the riots, it can also be postulated that civil unrest generated further interest in Robinson’s content, which in turn enhanced his opportunity to increase his platform and further fuel the protests. This finding adds nuance to the otherwise well-founded concern that “online false information contributes to political polarisation [via] echo chambers and filter bubbles, in which political opinions and prejudices are fostered through insular engagement” (Coe, 2023, p. 224; see also Dowdeswell & Goltz, 2020). Indeed, social upheaval might expand the reach of controversial commentators beyond their usual demographic.

Robinson’s ability to overcome the usual presumptions about algorithmic filtering, facilitated by Musk’s decisions as owner, might have been further strengthened by the direct support Musk often offers him as user. For instance, when Robinson accused the UK Prime Minister of labelling everyone upset about the Southport murders as “thugs,” Musk responded with two exclamation marks (Spring, 2024b). Musk is known to reply to or repost other X users’ content with emojis, exclamation marks or one-word comments (Maxwell, 2023), and while this might be because he genuinely finds associated content “cool” or “true,” the fact this creates a link between the content and Musk’s 200 million X followers cannot be ignored. Nor can the New York Times’ reporting that X’s algorithms prioritise its owner’s content (Conger et al., 2024; Schiffer & Newton, 2023). Musk’s interventions during the riots, however, were not limited to boosting the reach of Robinson’s posts. When conservative commentator Ashley St Clair argued on X that the UK riots were “the effects of mass migration and open borders” (St Clair, n.d.), Musk responded with the words “civil war is inevitable” (Musk, n.d.-c). This received 10 million views and 33,400 “likes” (Musk, n.d.-c). Downing Street’s response not only called Musk’s comments “unjustified” but also implied a link between them and the riots: “anyone who is whipping up violence online will face the full force of the law” (Brown & Culbertson, 2024).

Musk, however, was not deterred. When Starmer publicly pledged protection for the UK’s Muslim communities during the protests, Musk shared a video of “armed ‘Muslim patrol’ members…looking for white right-wingers to attack” asking: “Why aren’t all communities protected in Britain @keirstarmer?” (Musk, n.d.-d). Later, he used X to draw attention to riot-related arrests, particularly those resulting from online activity. His query, posted while the riots were ongoing, as to whether this was "Britain or the Soviet Union" was viewed 48 million times (Musk, n.d.-e). His post "#TwoTierKeir," also uploaded during the riots, implying that under the Labour government Muslims received preferential treatment over the white population, received 7.2 million views, 86,000 "likes," and 15,000 reposts (Musk, n.d.-f).

Musk's accusations of anti-white racial profiling continued after the riots. When Safeguarding Minister Jess Phillips opted to encourage a locally-led, rather than national, inquiry into grooming gangs in Oldham, Musk reposted a tweet from another X user, which accused Phillips of "refus[ing] an inquiry into [M]uslim grooming gangs," and asked "someone [to] tell her she is supposed to be safeguarding children not the gangs." Musk added his own view that "Jess Phillips is a rape genocide apologist." This post garnered 43.6 million views, 240,000 "likes," and 51,000 reposts (Musk, n.d.-g). In a separate post, viewed 60.2 million times and shared 46,000 times, Musk accused Starmer of being "complicit in the RAPE OF BRITAIN when he was head of the Crown Prosecution for 6 years. Starmer must go and he must face charges for his complicity in the worst mass crime in the history of Britain" (Musk, n.d.-h).

On the one hand, while extreme, it could feasibly be argued that Musk's posts, at least in relation to Phillips, were political speech and reflected his views on the decision to focus on a locally-led inquiry. On the other, Musk's unparalleled online reach, combined with the attention his posts inevitably receive from the traditional UK media (Sigsworth, 2025), arguably places him in a position of responsibility because there is a greater chance that his online speech could encourage digital and real-world hate crime. Indeed, Phillips reported that "threats to her own safety had gone up since [Musk's] social media posts" (BBC News, 2025b; see also Fenwick & Coe, 2025). A 39-year-old man was later convicted of sending malicious communications "saturated in hate and intolerance" to Phillips under s.1 of the Malicious Communications Act 1988 (Malicious Communications Act, 1988; PA Media, 2025). District Judge Smith stated that his email had caused the Safeguarding Minister "great distress" and that she "was concerned for [Bennett's] potential to escalate or encourage others for violence against her, having in mind the murder of her colleague Jo Cox" (PA Media, 2025). The Crown Prosecution Service noted that Bennett sent his email to Phillips "one day after Musk said the MP 'deserves to be in prison' for denying requests to the Home Office for a public inquiry into child sexual exploitation in Oldham" (PA Media, 2025).

As the above examples demonstrate, Musk stands in a position of responsibility at the nexus of online speech and real-world unrest. While he cannot be blamed for triggering the UK's summer 2024 riots, which were caused by a number of factors, his decisions to re-instate individuals who spread misinformation and/or emerged as figureheads of civil unrest, alongside his influence over X's content moderation and algorithm policies, and his personal choices as X's most high-profile user (Buchholz, 2025) mean he is uniquely positioned to amplify his chosen narratives. Indeed, the extent of his online reach means that he could claim to be a populist leader even in situations where he simply places a couple of exclamation marks over the posts of far-right activists. Crucially, as the next section underscores, this capacity is further buttressed by his offline influence and societal power.

2. Amplifying hate through societal power: Musk as a populist leader

The use of social media to express strong political views is hardly news. Indeed, at its advent, it was hoped that its transformation of the internet consumer into participant would unlock the web's democratising potential (Obar & Wildman, 2015; Ceron, 2017). Ordinary people could now disseminate their political views to the world at large. Not only was there no need for parliamentary or media intermediaries but these estates could now be held to account by this "fifth estate" of power (Dubois & Dutton, 2014). Yet further research has revealed "the attention economy of social media [to be] highly unequal" in practice (Ojala et al., 2018). As section one highlights, Musk provides the paradigm example. The online speech of a social media user, who also happens to be its owner, who possesses over 200 million followers, and who can influence algorithmic prioritisation is clearly going to have greater reach than the average citizen. These things alone enable Musk to operate as a populist leader and not just in the UK. It is well-documented that Musk has chosen to platform populist political parties in European elections. His tweet claiming that "only the AfD can save Germany" – sections of which have been found to be right-wing extremist by German authorities (Parker, 2025) – has been viewed 52.5 million times, liked 147,000 times, and shared 29,000 times (Musk, n.d.-i). Crucially, Nenno and Lorenz-Speen document a spike in the online reach of Alternative für Deutschland co-chair Alice Weidel around the time of Musk's post (Nenno & Lorenz-Speen, 2025). Importantly, they also outline that Musk's reach extends beyond online activity. His interview endorsing Weidel in Die Welt, combined with a related exchange of X posts between Musk and Weidel, similarly coincided with an upturn in Weidel's X followers and the number of views her posts received. Interestingly, from the data available to them(Nenno & Lorenz-Speen, 2025; House of Lords Communications and Digital Committee, 2024; Dowdeswell & Goltz, 2020), Nenno and Lorenz-Speen found "no evidence of an adjustment to the algorithm in Weidel's favour," concluding that while such changes might have been subtle, the more obvious contributor might have been "that Musk's statements on X are particularly visible and a retweet from him can lead to a shift in discourse" (Nenno & Lorenz-Speen, 2025). Accordingly, it is significant that Musk handed Weidel extended access to his user-base via a 74-minute livestream from his X account, during which he interviewed the AfD co-leader and endorsed the party (Anon, 2025)

Other German politicians view Musk's interventions as democratically problematic. Dirk Wiese of the centre-left SPD party alleged that Musk "use[s] his influence for deliberate provocations, misinformation, and the spread of populist narratives" (Watling & Hagopian, 2025). Of course, such statements are expected from politicians whose parties are in electoral competition with the AfD. Nevertheless, that constitutional actors feel it necessary to comment points to an understanding of Musk's societal influence, which in turn points to his potential to lead populist narratives. Accordingly, having acknowledged Musk's online reach above and in section one, this section now turns to examining the reactions of established constitutional actors to Musk's political interventions. It argues that these responses, reluctant or otherwise, risk cementing rather than counteracting Musk's messaging. This raises imperative questions about unchecked private power, which escapes constitutional checks and balances whilst drawing on those same safeguards to perpetuate populist narratives and further commercial dominance.

Certainly in the UK, various high-profile political figures have considered it wise to work with, rather than against, Musk. Even after he had been accused of fanning the flames of the 2024 riots and at a time when he had just labelled the UK "a Stalinist state" (Musk, n.d.-j), Labour party grandee Peter Mandelson publicly called on Keir Starmer to ease tensions with the X owner for unavoidable pragmatic reasons. Mandelson declared Musk "a sort of technological, industrial and commercial phenomenon [who]…it would be very unwise in my view for Britain to ignore… You've got to get over it…he's got to be re-introduced to the British government" (Maddox, 2024). While Starmer did not opt for this approach at that time, he did still recommend Mandelson for ambassador to Washington, just as Musk was heading up DOGE (Keate, 2025). This is perhaps only a small illustration of a wider problem identified by Alemanno and Veraldi, who argue that Musk's "multi-industry influence gives rise to profound questions about the limits of individual influence and power accumulation in a complex geopolitical landscape" (Alemanno & Veraldi, 2025, p. 11).

Beyond these subtle government responses, the constant interest in Musk by another established constitutional actor – the traditional media – risks diverting attention from pertinent issues towards more sensationalist matters, which in turn can impact upon the execution of constitutionally essential accountability. For instance, when the UK Prime Minister delivered a speech on his Government's strategy for reducing waiting lists in the country's National Health Service, the press nonetheless asked him to comment on an X poll run by Musk as to whether "America should liberate the people of Britain from their tyrannical government?" (Musk, n.d.-k). Starmer refused to comment and so was asked about Musk again (Cooke, 2025). This time the question concerned the risks posed by Musk's online posts to Jess Phillips. Despite asserting that "on the question of Elon Musk, I think most people are more interested in what is going to happen with the NHS quite frankly," Starmer was forced to provide a comment when pressed on the matter yet again: "those that are spreading lies and misinformation as far and wide as possible are not interested in victims, they're interested in themselves. They are cheerleading Tommy Robinson, a man who went to prison for nearly collapsing a grooming case" (Cooke, 2025).

Critically, although the media reported on Labour's waiting list reduction strategy (Crerar & Campbell, 2025), any criticism of its approach (Gregory, 2025) quickly became chip paper in the face of continued press and parliamentary attention to the question of "Muslim grooming gangs" as pushed by Musk. This despite the fact that the chair of an independent inquiry into child sex abuse, which had taken place as recently as 2022 (Independent Inquiry into Child Sexual Abuse, 2022), had publicly stated that a further inquiry would "certainly cause delays" to the implementation of its recommendations (Mercer & Zeffman, 2025). "Galvanised" by Musk (Rigby, 2025a), the Conservative Opposition tabled an amendment to the Children's Wellbeing and Schools Bill to cover the issue despite having rejected similar calls when in power themselves (Anon, 2025b). Importantly, this amendment would not have brought an inquiry into existence but would incidentally have "killed the government's legislation, the aim of which [was] to reform…the children's care system and raise educational standards" (Brown, 2025).

Ultimately, given Labour's large parliamentary majority, the amendment was defeated. The media underscored, however, that it was "largely symbolic"; its role being to contribute to broader political pressure on Government "following days of headlines after comments by Elon Musk" (Brown, 2025). Subsequently, although the UK Government resisted a full U-turn, it did announce a "national review" into grooming gang evidence (Whannel & Thomas, 2025), alongside funding for five Government-backed local inquiries into child sexual abuse. Perhaps in light of Musk's personal attacks on Jess Phillips over the issue, Government Ministers "utterly rejected" any notion that his interventions had triggered the policy shift (Francis, 2025). The BBC's chief political correspondent, Henry Zeffman, however, begged to differ: "we should not lose sight of the fact that there is no chance [Home Secretary Yvette] Cooper would have been standing up in the House of Commons [announcing a national review] today if Elon Musk hadn't taken an interest in it…and that remains an extraordinary…fact that I just think we need to continue to reflect on because it could well be heralding…a whole new way in which our information and political systems interact" (Rigby, 2025a).

Indeed, the above illustrations of Musk's political influence suggest that even if it is sometimes begrudging, established constitutional players do not always counteract but rather cement or even amplify his narratives. The next question is why? This might simply be a case of political strategizing, as the UK Conservatives' own U-turn on grooming gangs when going from Government to Opposition implies. This would nonetheless emphasise Veraldi and Alemanno's argument that Musk is not only an "economic behemoth but now wields disproportionate influence over the information ecosystem, civic discourse and the shaping of political agendas" (Veraldi & Alemanno, 2025, p. 155), enabling "personal bias to shape…narratives" (Veraldi & Alemanno, 2025, p. 155). Crucially, such developments point to a wider shift as regards the position of private power in contemporary constitutional landscapes. As Koulierkis highlights, "there is a striking yet neglected interplay between immense economic power and high politics" (Koulierkis, 2025, p. 109).

How then has this interplay come about? Of course, the answer seems intrinsically obvious. Musk is the world's richest person and the on-again-off-again right-hand man of US President Trump (Drenon, 2025; Cai, 2025). Yet discussion of the relationship between economic might and the constitutional impacts born from growing populism and political polarisation is arguably rather neglected in the UK constitutional literature, which despite being heavily focused on political components tends to restrict political analyses to parliamentary figures and practices (Reynolds, 2025). Knox, however, has convincingly advanced that what constitutes constitutional activities, impacts and/or actors is shaped by the "legislative and material conditions" of the time (Knox, 2022, pp. 324, 339). He draws clear links, for instance, between the socio-democratic model pursued in the UK during the post-war consensus, the legislative interventions of the Attlee Government, and the strengthening of trade unions' societal power during the 1970s (Knox, 2022).

In a similar way, we might look to the neo-liberal model, with its promotion of the individual, of capital accumulation, and free markets, as a driver of Musk's societal heft and therefore as an enabler of his de facto populist leadership. Von Thun postulates:

> [W]hen it comes to the extreme concentration of wealth and power in the hands of a small number of American oligarchs and corporations, the US government itself naturally bears most of the blame. By slashing taxes on high incomes and wealth, weakening labour unions and crippling competition enforcement under President Reagan onwards, the US government denuded itself of some of the most powerful tools for tackling plutocracy. (von Thun, 2025, p. 81)

Furthermore, he posits, "Europe imported neoliberal thinking from the US, which placed market efficiency above all other considerations – including democratic and societal resilience" (von Thun, 2025). A key aspect of neo-liberalism has been to bring private sector innovation into public service delivery. When these innovators also run the social media platforms that proliferate hate speech, governments find themselves hamstrung. Thus, "tens of thousands of Europeans are dependent on [Musk's] Starlink for Internet connectivity," while Musk's SpaceX rockets are being "used by the EU to launch satellites and telescopes" (Alemanno & Veraldi, 2025, p. 12). As evidenced by his threats to remove support, Starlink and SpaceX are pivotal to the Ukrainian war effort (Alemanno & Veraldi, 2025). More broadly, as von Thun highlights, Europe is dependent on Amazon, Microsoft, and Google for cloud computing infrastructure, and on Google for search and digital advertising, all of which have come under fire for the contributions their technology can make to the spread of dis/misinformation (von Thun, 2025). The result, perhaps, is that despite Keir Starmer's clear irritation when it came to Musk's intervention on grooming gangs (Rigby, 2025b), "no political European leader seems capable or willing to oppose his frontal attack on the European continent" (Alemanno & Veraldi, 2025, p. 12).

For Golia, the problem, however, is not just neo-liberalism's effects on the constitutional environment but the presumptions of liberal constitutionalism itself and, in particular, its maintenance of a public/private divide (Golia, 2024). As Grimm highlights, private actors have progressively gained a greater share of public power without being integrated into constitutional accountability frameworks designed to constrain the coercive powers of the State (Grimm, 2010). In the context of online speech, the literature's focus has been on State exploitation of this divide via mechanisms that utilise private platforms' terms of use, rather than legislation, to police speech because, as private actors, social media platforms are not subject to formal human rights obligations (Alkiviadou, 2024). These concerns remain legitimate. However, an additional consideration must now be the private co-opting of this divide, and of the individual right to freedom of expression, by tech magnates in ways which permit the perpetuation of online content, including hate speech and mis/disinformation, which entrenches their commercial position.

Musk has declared himself a "free speech absolutist," claiming that he did not buy Twitter in order to "make more money. I did it to help humanity" (Anon, 2022b). Prior to his takeover of the platform he was highly critical of its content moderation policies, using his own account to launch a poll asking "what should be done?" given that "Twitter serves as the de facto public town square [and] failing to adhere to free speech principles undermines democracy" (Milmo, 2022). It is entirely possible that Musk is genuine when he communicates this belief, even if he has been accused of hypocrisy in this regard (Counts, 2023). His decision to reactivate the account of previously banned users and to weaken the platform's content moderation policies when he took over could well reflect this. After all, such an approach reflects established arguments about the marketplace of ideas made in the literature and case law. John Stuart Mill argued that the "foundations and reasoning upon which opinions are based must be continually tested and, as a result, the acceptance of alternative views by others, and ultimately the reliable discovery of truth, must derive from effective persuasion rather than coercion" (Coe, 2023, p. 244). In the US case of Abrams v United States, Justice Holmes concluded that "the best test of truth is the power of the thought to get itself accepted in the competition of the market" (Abrams v. United States, 1919, p. 630). Meanwhile in Packingham v North Carolina, Justice Kennedy referred to the "internet [as] the modern public square" (Packingham v. North Carolina, 2017, p. 1737). The admittedly much-contested s.230 Communication Decency Act, which protects platforms from responsibility for content posted on their sites by users, reflects the US view that free speech requires heightened protection from Government interference; itself underscored by the core tenets of liberal constitutionalism, which seek to safeguard individual liberty from State power.

Yet, as Coe has argued, in practice, social media undermines the marketplace of ideas and the Habermasian concept of discourse in a rational public sphere because it "proliferates a huge amount of information that is poorly researched or simply untrue, yet has the potential to, and very often does emerge as the dominant 'view' regardless of the detrimental impact this may have on individuals or society" (Coe, 2023, p. 225). Despite this, following the election of President Trump, Meta's owner Mark Zuckerberg announced changes to its moderation policies, including the elimination of fact-checkers in order, he claimed, to prioritise freedom of expression, which, he implied, had been undermined as a result of government pressure (Bayer, 2025).

While Zuckerberg's reasons might be genuine, it is interesting that these changes coincided with the arrival of President Trump, increased focus by the US administration on alleged free speech restrictions and censorship in Europe (Bourne & Whannel, 2025), and stronger European attempts to regulate digital platforms, for instance via the UK's Online Safety Act 2023 and the EU's Digital Services Act. These are understood to introduce commercial risks for the tech bros. In other words, as Davies and Cogen identify, "having amassed huge amounts of wealth by placing themselves at strategic chokepoints in the economy, [tech oligarchs] have shown themselves to be adept at converting economic power into political and cultural power and then back again" (Davies & Cogen, 2024, p. 13). As the UK's House of Commons Committee on Digital, Culture, Media and Sport observes:

The prevalence of disinformation online must be understood within the business context of tech companies. Tech companies generate revenue primarily through advertising targeted at users based on observed or perceived tastes and preferences, which is maximised by increasing the user base, data collection, average user time and user personalisation. We know that novelty and fear (along with anger and disgust) are factors which drive "engagement" with social media posts; that in turn pushes posts with these features further up users' newsfeeds – this is one reason why false news can travel so fast… The more people engage with conspiracy theories…the more platforms are incentivised to continue surfacing content, which theoretically encourages users to continue using the platform so that more data can be collected and more adverts can be displayed. (House of Commons Digital, Culture, Media and Sport Committee, 2020, p. 13; see also Mattiuzzo & Nicolini de Morais, 2024)

Similarly, Ong and Toh argue that platforms' digital dominance "has enabled them to pursue profit-driven operational strategies that promote extremist viewpoints, polarise public opinion, facilitate dissemination of information and jeopardise physical and mental well-being" (Ong & Toh, 2023, p. 563). Van der Kerkhof concedes that Musk has "delivered on his promise of free speech absolutism and has indeed created a virtually lawless public square" but notes that "[i]n that promise he downplays his role in amplifying certain voices" within it (van der Kerkhof, 2025, p. 96). Once again, the public/private divide in liberal constitutionalism and associated human rights law reinforce Musk's prerogatives in this regard. As a report to the European Parliament confirmed, under "US free speech doctrine, private companies are not obliged to respect free speech and are [generally] allowed to act as they like against the free speech of others" (Bayer et al., 2021, p. 84). This is reinforced by the protections liberal constitutional models afford to private property. As Cowls et al. observe, "platform owners are able to engage in content moderation through their rights of ownership" (Cowls et al., 2020, p. 3). Even in Europe, where meaningful attempts have recently been made to regulate digital services – triggering the ire of the likes of Musk and Zuckerberg – the legitimate need to balance curbs on hate speech with the protection of free expression have largely left the world's private "digital governors" in control despite the fact that "informational capitalism remains at the core of their business model" (Golia, 2024, pp. 510–511). An assessment of whether the UK's recent Online Safety Act 2023 could do anything to address Musk's posts demonstrates this.

3. Testing the Online Safety Act 2023 against Musk's X posts

The UK Government has touted its Online Safety Act as legislation that makes the UK "the safest place in the world to be online" (HM Government, 2023). The OSA has, however, come under fire for "fail[ing] to further democratic ends while also failing to curb anti-democratic harms" (Fenwick & Coe, 2025, p. 87), largely because with the exception of some new criminal offences and limited removal requirements, the Act largely relies on platforms' own terms of service, despite its original objectives of bringing an end to self-regulation. These accusations seem to be made out when the Act is tested against Musk's tweets. Section 181 OSA makes it an offence for an individual to send a message that conveys the threat of death or serious harm with the intention, or being reckless, as to whether the individual encountering the message would fear that that threat would be carried out. However, while Jess Phillips reported feeling increasingly threatened following Musk's online posts), it seems highly unlikely that his calls for her to be jailed or his accusations that she was a "rape genocide apologist" would meet the high threshold set for this offence.

Likewise, these interventions seem unlikely to trigger the OSA's "false communications offence" under s.179. This has a high mens rea threshold, requiring the defendant to know that the information he/she was sending was false and to intend that information to cause non-trivial harm to "a likely audience." There must also be an absence of "reasonable excuse." The extreme language used by Musk aside, it seems unlikely that political comment on a decision not to launch a national inquiry into grooming gangs would qualify. The same conclusions seem likely in relation to Musk's accusations against "TwoTierKeir" and his related sharing of a video, claiming to be of the aftermath of a Muslim stabbing of white protestors. It would be hard to establish that Musk knew the information attached to the video was false and that he intentionally shared it anyway. In any case, ss.179 and 181 only have territorial application if the act is done by an individual habitually resident in England and Wales or Northern Ireland (Online Safety Act 2023, s.185). While it is perhaps more significant to ask whether charges of this nature would ever realistically be brought against Musk, in any case, given the increasing public dependence on his private endeavours, the high thresholds attached to these offences also apply to the ordinary user.

Section 127(1) of the Communications Act 2003 does render it an offence to "send by means of a public electronic communications network a message…that is grossly offensive." As Fenwick and Coe acknowledge, this provision has been "deployed on a number of occasions in relation to online threats to female MPs" but is not categorised by the OSA as "priority illegal content" that would trigger an obligation on social media platforms to remove it (Fenwick & Coe, 2025, p. 87). When it comes to "legal but harmful content," the OSA simply imposes duties on large service providers like X to impose their own terms and conditions. The Government conceded that "legal but harmful" content might include "racist, misogynistic" posts, simply saying that "the largest and riskiest companies will need to set out clearly in terms and conditions what harmful material is allowed on their site." The need to "protect our core democratic rights, including freedom of expression… [meant that] it would be inappropriate for the Government to require companies to remove legal content accessed by adults" (HMG, 2022). As Fenwick and Coe underscore, "leaving it up to the services to determine the content that is covered in their service terms, and relying on them to apply it consistently, makes those services de facto arbiters of free speech…some speech of political value might be removed while, conversely, some harmful expression of anti-democratic tendency might not be" (Fenwick & Coe, 2025, pp. 125–126). Arguably, the recognition in X's terms of service that "sometimes it might be in the public interest to allow people to view posts that would otherwise violate our policies" (X, n.d.-b) will work to the benefit of Musk and other high-profile users.

The crux of the matter is that free speech concerns are legitimate. It might, therefore, be almost impossible to square the free speech circle when it comes to regulating out hate speech. Coe, for instance, "is not convinced that the OSA, and regulation generally, is the appropriate way to meet the challenge we face, as it is not the radical panacea for protecting us from online harms that it has been presented as" (Coe, 2023, p. 240). Instead, he points, inter alia, to proper funding for regulators and the improvement of digital education to build digital resilience as important first steps (Coe, 2023, p. 240; see also Counts, 2023, p. 499). While these will be crucial contributors to tackling the problem, these solutions still approach its causes as though they were constrained within the digital realm. As Golia highlights, "digital tech and globalisation have not created but made more visible and urgent, questions left unaddressed by state-centred liberal constitutionalism" (Golia, 2024, pp. 510–511). In particular, we might look to the public/private divide it reinforces and to neo-liberalism's promotion of private power. Populism, Loughlin emphasises, works from within and is a symptom of the degradation, rather than overthrowing, of our constitutional democracies (Loughlin, 2019). This has been caused, inter alia, by expansions in global trade, global communications and global investment and the weakening of national legislatures as compared with executives and non-majoritarian institutions, which have removed citizens' participation from things that affect their lives. Neo-liberal approaches have created rules to protect the market whilst overlooking the preservation of democratic self-determination. Resulting "extreme wealth imbalances raise constitutional issues by undermining the common feeling that sustains republican government" (Loughlin, 2019). In this context, it pays for billionaire populists to divert the attention of disempowered populations away from themselves and towards other groups, while internet conspiracies make those same populations "feel empowered" (Wright, 2024, oral evidence, HC 175). The real challenge for the modern state, therefore, is to restore citizen empowerment, thereby stripping hate speech of its seductive power.

References

Abrams v. United States, 250 U.S. 616 (1919).

Alemanno, A., & Veraldi, J. D. (Eds.). (2025). *Musk, power and the EU: Can EU law tackle the challenges of unchecked plutocracy?* Verfassungsbooks.

Alkiviadou, N. (2024). Platform liability, hate speech and the fundamental right to free speech. *Information and Communications Technology Law*, 34(2), 1–11.

Anon. (2022a). Elon Musk's Twitter dissolves Trust and Safety Council – just days after its members speak out. *Sky News*. https://news.sky.com/story/elon-musks-twitter-dissolves-trust-and-safety-council-just-days-after-its-members-speak-out-12767148

Anon. (2022b). Elon Musk claims he is buying Twitter to help humanity. *BBC News Online*. https://www.bbc.co.uk/news/business-63408384

Anon. (2025a). Jess Phillips does not rule out new national inquiry into grooming gangs. *Sky News*. https://news.sky.com/story/jess-phillips-does-not-rule-out-new-national-inquiry-into-grooming-gangs-13285393

Anon. (2025b). Elon Musk livestreams chat with German far-right leader. *Newsweek*. https://www.newsweek.com/elon-musk-alice-weidel-x-livestream-2012828

Bayer, J. (2025). What Big Tech Brothers' state capture means for the European Union. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU* *law tackle the challenges of unchecked plutocracy?* (pp. 32–48). Verfassungsbooks.

Bayer, J., Bárd, P., Szakács, J., Alemanno, A., & Uszkiewicz, E. (2021). *Disinformation and propaganda: Impact on the functioning of the rule of law and democratic processes in the EU and its Member States – 2021 update*. European Parliament (INGE Committee).

BBC News. (2025a, February 27). *Who is Andrew Tate? The self-proclaimed misogynist influencer*. https://www.bbc.co.uk/news/uk-64125045

BBC News. (2025b, January 7). *Musk's "disinformation" endangering me, says Phillips*. https://www.bbc.co.uk/news/articles/cn7r0pzz57vo

Bourne, V., & Whannel, K. (2025, August 13). US says UK human rights have worsened in past year. *BBC News Online*. https://www.bbc.co.uk/news/articles/cqjyeeke7qko

Brewster, T. (2024, January 10). Musk's X fired 80% of engineers working on trust and safety, Australian Government says. *Forbes*.

Browne, O. (2024, August 8). Andrew Tate admits he was wrong about Southport suspect but claims he is "not the reason the riots started." *AOL Online / The Independent.*

Brown, F. (2025, January 8). MPs vote against new national inquiry into grooming gangs. *Sky News*. https://news.sky.com/story/mps-vote-against-new-national-inquiry-into-grooming-gangs-13285629

Brown, F., & Culbertson, A. (2024, August 5). Elon Musk hits back at Sir Keir Starmer after "civil war" comments dismissed. *Sky News*. https://news.sky.com/story/pm-warns-anyone-whipping-up-online-violence-to-face-full-force-of-law-after-elon-musk-civil-war-comments-13191316

Buchholz, K. (2025, November 28). X followers: All about Elon? *Statista*. https://www.statista.com/chart/35367/accounts-with-the-most-followers-on-x/

Cai, S. (2025, October 27). Musk back in Trump's good graces after summer of public feuding. *Politico*. https://www.politico.eu/article/us-elon-musk-back-in-donald-trump-good-graces-after-summer-of-public-feuding/

Centre for Countering Digital Hate. (2024, August 7). *Musk's X helps Tommy Robinson rack up 434 million views during UK riots*. https://counterhate.com/research/musks-x-helps-tommy-robinson-rack-up-434-million-views-during-uk-riots/

Ceron, A. (2017). *Social media and political accountability.* Springer.

Coe, P. (2023). Tackling online false information in the United Kingdom: The Online Safety Act 2023 and its disconnection from free speech law and theory. *Journal of Media Law*, 15(2), 213–242.

Community Note on X. (n.d.). Note attached to post about a "police station" being destroyed. *X* (formerly Twitter).

Conger, K., et al. (2024, November 3). How Elon Musk's own account dominates X. *The New York Times*.

Cooke, M. (2025, January 8). A history of Elon Musk and Keir Starmer's relationship, from the Southport riots to grooming gang claims. *The Independent*.

Cowls, J., et al. (2020, November 3). *Freedom of expression in the digital public sphere*. Tilburg University. https://repository.tilburguniversity.edu/server/api/core/bitstreams/326f03b7-6b24-46db-8b68-f12598a1e92e/content

Counts, A. (2023, September 14). Elon Musk is a "free speech absolutist", except at work. *Bloomberg UK*. https://www.bloomberg.com/news/newsletters/2023-09-14/elon-musk-says-he-s-pro-free-speech-but-fired-twitter-staff-for-comments

Crerar, P., & Campbell, D. (2025, January 10). Streeting defends NHS use of private sector but says it must "pull its weight." *The Guardian*.

Cumming, E. (2024, October 30). How the Southport terror charge is undermining trust in the police. *The Telegraph*.

Davies, T., & Cogen, S. (2024). Democracy or domination: The role of competition law in the face of oligarchy. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU law tackle the challenges of unchecked plutocracy?* (pp. 55–76). Verfassungsbooks.

Drenon, M. (2025, June 7). Trump says relationship with Musk is over. *BBC News Online*. https://www.bbc.co.uk/news/articles/c9wg240q0plo

Dubois, E., & Dutton, W. H. (2014). Empowering citizens of the internet age: The role of a fifth estate. In M. Graham & W. H. Dutton (Eds.), *Society and the internet: How networks of information and communication are changing our lives*. Oxford University Press.

Dowdeswell, T., & Goltz, N. (2020). The clash of empires: Regulating technological threats to civil society. *Information and Communications Technology Law*, 29(2), 194–217.

Fenwick, H., & Coe, P. (2025). The OSA: Fostering democratic participation while combatting anti-democratic harms? *Northern Ireland Legal Quarterly*, 76(AD1), 83–135.

Forbes. (n.d.). The world's real-time billionaires. https://www.forbes.com/real-time-billionaires/

Francis, S. (2025, January 17). Nandy denies Musk prompted grooming gangs inquiry. BBC News Online. https://www.bbc.co.uk/news/articles/c05ll7r9vd9o

Full Fact. (n.d.). *Two stabbings in Stoke claim is false*. https://fullfact.org/online/two-stabbings-stoke-false/

Golia, A. (2024). Critique of digital constitutionalism: Deconstruction and reconstruction from a societal perspective. *Global Constitutionalism*, 13(3), 488–518.

Griffin, A. (2024, September 6). How a few Twitter posts on Elon Musk's X helped fan the flames of unrest and rioting across the UK. *The Independent*.

Grimm, D. (2010). The achievement of constitutionalism and its prospects in a changed world. In P. Dobner & M. Loughlin (Eds.), *The twilight of constitutionalism?*. Oxford University Press.

Gregory, A. (2025, January 3). Ministers plan biggest shake-up of adult social care in England for decades. *The Guardian*.

HM Government. (2023, October 26). *UK children and adults to be safer online as world-leading bill becomes law* [Press release]. https://www.gov.uk/government/news/uk-children-and-adults-to-be-safer-online-as-world-leading-bill-becomes-law

HMG. (2022, March 22). *The draft Online Safety Bill and the legal but harmful debate: Government response to the DCMS Committee 8th report* (5th Special Report of Session 2021–22, HC 1221).

House of Commons Digital, Culture, Media and Sport Committee. (2020). *Misinformation in the Covid-19 infodemic* (Second Report of Session 2019–21, HC 234).

House of Commons Digital, Culture, Media and Sport Committee. (2024). *Trusted voices* (Sixth Report of Session 2023–24, HC 175).

House of Lords Communications and Digital Committee. (2024). *The future of news* (1st Report of Session 2024–25, HL Paper 39).

Independent Inquiry into Child Sexual Abuse. (2022). *Report of the Independent Inquiry into Child Sexual Abuse* (HC 720). https://www.iicsa.org.uk/document/report-independent-inquiry-child-sexual-abuse-october-2022-0.html

Keate, N. (2025, September 11). Peter Mandelson sacked as British ambassador to US over Epstein friendship. *Politico*. https://www.politico.eu/article/peter-mandelson-resigns-as-british-ambassador-to-us-over-jeffrey-epstein-friendship/

Knox, R. (2022). Neo-liberalism, labour law and New Labour's turn to constitutionalism. In M. Gordon & A. Tucker (Eds.), *The New Labour constitution: Twenty years on* (ch. 15, pp. 324–339). Hart.

Koulierkis, E. (2025). Empowering citizens against technological corporations. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU law* *tackle the challenges of unchecked plutocracy?* (pp. 109–128). Verfassungsbooks.

Lindsay, J. (2018, August 1). Why did Tommy Robinson change his name from real name Stephen Yaxley-Lennon? *Metro*.

Loughlin, M. (2019). The contemporary crisis of constitutional democracy. *Oxford Journal of Legal Studies*, 39(2), 435–454.

Maddox, B. (2024, November 19). Peter Mandelson urges Starmer to recruit Nigel Farage to help woo Donald Trump and Elon Musk. *The Independent*.

Malicious Communications Act 1988, c. 27 (UK).

Marshall, J. (2024, June 5). Belfast violence: What happened at the weekend? *BBC News Online*. https://www.bbc.co.uk/news/articles/c9wjjr7wq12o

Mattiuzzo, M., & Nicolini de Morais, J. C. (2024). Social media and deceptive patterns: A way forward for anti-trust enforcement. *Journal of Competition Law and Economics*, 20(4), 343–383.

Maxwell, T. (2023, January 25). Elon Musk is a prolific Twitter replier. From population collapse to laughing emojis, here are his top replies. *Business Insider*. https://www.businessinsider.com/elon-musk-twitter-replies-most-common-population-collapse-2023-1

Mercer, D., & Zeffman, H. (2025, January 7). Victims want action, child abuse inquiry chair says. *BBC News Online*. https://www.bbc.co.uk/news/articles/cp836w074gko

Merseyside Police. (2024, July). *Three girls tragically killed in Southport named.* https://www.merseyside.police.uk/news/merseyside/news/2024/july/three-girls-tragically-killed-in-southport-named/

Milmo, D. (2022, April 14). How "free speech absolutist" Elon Musk would transform Twitter. *The Guardian*.

Milmo, D., & Quinn, B. (2024, July 31). How false online claims about Southport knife attack spread so quickly. *The Guardian*.

Moynihan, D. (2025). The death of USAID: How Elon Musk and Donald Trump ended America's foreign aid agency. *Public Administration and Development*, 45(4), 327–331.

Musk, E. (n.d.-a). What is happening in the UK?! [Tweet]. *X*. https://x.com/elonmusk/status/1860771593682817438

Musk, E. (n.d.-b). [Tweet about forming a new American political party]. *X*. https://x.com/elonmusk/status/1930685402631053403

Musk, E. (n.d.-c). Civil war is inevitable [Reply to Ashley St Clair]. *X*. https://x.com/elonmusk/status/1819933223536742771

Musk, E. (n.d.-d). Why aren't all communities protected in Britain @keirstarmer? [Tweet with "Muslim patrol" video]. *X.* https://x.com/elonmusk/status/1820790297233592361

Musk, E. (n.d.-e). Is this Britain or the Soviet Union? [Tweet]. *X.* https://x.com/elonmusk/status/1820784815815160260

Musk, E. (n.d.-f). #TwoTierKeir [Tweet]. *X.* https://x.com/elonmusk/status/1820805621534400786

Musk, E. (n.d.-g). Jess Phillips is a rape genocide apologist [Tweet]. *X.* https://x.com/elonmusk/status/1875145167633887358

Musk, E. (n.d.-h). Starmer complicit in the RAPE OF BRITAIN… [Tweet]. *X.* https://x.com/elonmusk/status/1875150194909823085

Musk, E. (n.d.-i). Only the AfD can save Germany [Tweet]. *X.* https://x.com/elonmusk/status/1869986946031988780

Musk, E. (n.d.-j). The UK is a Stalinist state [Tweet]. *X.* https://x.com/elonmusk/status/1858521463889825807

Musk, E. (n.d.-k). America should liberate the people of Britain from their tyrannical government? [Poll]. *X.* https://x.com/elonmusk/status/1876174862747930717

Nenno, S., & Lorenz-Speen, P. (2025, January 29). *Do Alice Weidel and the AfD benefit from Musk's attention on X?* Alexander von Humboldt Institute for Internet and Society – Digital Society Blog. https://www.hiig.de/en/musk-x-and-the-afd/

Obar, J. A., & Wildman, S. (2015). Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, 39(9), 745–750.

Ojala, M., et al. (2018). Networked publics as agents of accountability: Online interactions between citizens, the media and immigration officials during the European refugee crisis. *New Media & Society*, 21(2), 279–297.

Ong, B., & Toh, D. J. (2023). Digital dominance and social media platforms: Are competition authorities up to the task? *IIC – International Review of Intellectual Property and Competition Law*, 54, 527–572.

Online Safety Act 2023, c. [UK statute] (UK).

Ortutay, B., et al. (2022, November 29). Musk takes over Twitter, fires content moderation chief and now policing hate speech is his job. *Fortune*.

PA Media. (2025, February 18). Devon man jailed for sending "utterly deplorable" email to Jess Phillips MP. *The Guardian*.

Packingham v. North Carolina, 582 U.S. 98 (2017).

Parker, J. (2025, January 10). Musk interviews German far-right frontwoman.

BBC News Online. https://www.bbc.co.uk/news/articles/cr7errxp5jmo

Rawlinson, K. (2018, March 26). Tommy Robinson permanently banned from Twitter. *The Guardian*.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services (Digital Services Act), OJ L 277/1 (27.10.2022).

Reynolds, S. (2025). The high-profile private individual as constitutional actor. *King's Law Journal*, 36(1), 104–145.

Rigby, B. (2025a, January 9). Elon Musk's abuse of Jess Phillips has pushed real victims into game of political point scoring. *Sky News*. https://news.sky.com/story/elon-musks-abuse-of-jess-phillips-has-pushed-real-victims-into-game-of-political-point-scoring-13285656

Rigby, B. (2025b, January 7). New year, new Starmer? Why PM decided to finally take on Musk's "dangerous disinformation." *Sky News*. https://news.sky.com/story/new-year-new-starmer-why-pm-decided-to-finally-take-on-musks-dangerous-disinformation-13284732

Robinson, T. [@TRobinsonNewEra]. (n.d.-a). People need to rise up…our daughters are being butchered [Tweet]. *X*. https://x.com/TRobinsonNewEra/status/1818311713646682204

Robinson, T. [@TRobinsonNewEra]. (n.d.-b). The people revolt [Video of "police station"] [Tweet]. *X*. https://x.com/TRobinsonNewEra/status/1819472851797148087

Robinson, T. [@TRobinsonNewEra]. (n.d.-c). Report of two #EnoughisEnough protestors stabbed by Muslims [Tweet]. *X*. https://x.com/TRobinsonNewEra/status/1819747711052042580

Robinson, T. [@TRobinsonNewEra]. (n.d.-d). [Tweet sharing his own engagement statistics]. *X*. https://x.com/TRobinsonNewEra/status/1820678890727133566

Schiffer, Z., & Newton, C. (2023, February 15). Yes, Elon Musk created a special system for showing you his tweets first. *The Verge*. https://www.theverge.com/2023/2/14/23600358/elon-musk-tweets-algorithm-changes-twitter

Sigsworth, T. (2025, January 8). Elon Musk adds hypocrisy to his charge sheet against Jess Phillips. *The Telegraph*.

Sneed, T. (2025, March 30). Is DOGE actually an agency? The answer could have major ramifications. *CNN Politics*.

Spring, M. (2024a, July 31). Did social media fan the flames of riot in Southport? *BBC News Online*. https://www.bbc.co.uk/news/articles/cd1e8d7llg9o

Spring, M. (2024b, June 7). What is Elon Musk's game plan? *BBC News Online*.

https://www.bbc.co.uk/news/articles/cze5gd1jzkeo

St Clair, A. [@stclairashley]. (n.d.). The UK riots are the effects of mass migration and open borders [Tweet]. *X*. https://x.com/stclairashley/status/1819897308265214404

Statista. (n.d.-a). Twitter (now X) formerly banned accounts reinstated after Musk take-over: Number of followers. https://www.statista.com/statistics/1350470/twitter-formerly-banned-accounts-reinstated-musk-take-over-number-of-followers/

Unknown. (2024, August 7). Tommy Robinson's tweets from Cyprus viewed 50m times a day: The far-right activist – unbanned by Elon Musk – has had a surge in traffic since fleeing Britain. *The Times*.

van der Kerkhof, J. (2025). Musk, techbrocracy and free speech. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU law tackle the challenges of unchecked plutocracy?* (pp. 96–108). Verfassungsbooks.

Veraldi, J., & Alemanno, A. (2025). Does the EU have what it takes to counter American plutocratic power? A research agenda on power in the EU. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU law tackle the challenges of unchecked plutocracy?* (pp. 155–178). Verfassungsbooks.

von Thun, M. (2025). Europe vs the tech plutocrats: An existential battle for democracy and sovereignty. In A. Alemanno & J. D. Veraldi (Eds.), *Musk, power and the EU: Can EU law tackle the challenges of unchecked plutocracy?* (pp. 79–94). Verfassungsbooks.

Watling, T., & Hagopian, A. (2025, February 24). Elon Musk wants the far-right AfD to win the German election – here's how he became their champion. *The Independent*.

Whannel, K., & Thomas, E. (2025, January 16). Cooper announces inquiries into grooming gangs. *BBC News Online*. https://www.bbc.co.uk/news/articles/c9w5l4vxv2qo

Willis, R. (2020). Exploring the relationship between global Twitter campaigns and domestic law: Methodological challenges and solutions. *Information & Communications Technology Law*, 30(1), 3–16.

X. (n.d.-a). *About X Premium*. https://help.x.com/en/using-x/x-premium

X. (n.d.-b). *Public-interest exceptions.* https://help.x.com/en/rules-and-policies/public-interest

Part II

Populism and Digital Polarization: The Politics of Online Hate

Digital Populism and the Legitimization of Hate: Communication Strategies, Social Cohesion, and the Role of Online Platforms

Ibrahim Kurt, Society of European Scholars

1. Introduction

Around the world, populist politics have significantly increased in popularity in the twenty-first century. Populist politicians have used popular social worries, economic disparities, and political disenchantment to create compelling narratives of "the people" against "the elites" in a number of European countries, including Brazil, the United States, and Turkey. Populism portrays itself as a movement that empowers "ordinary citizens" in opposition to the establishment's indifference, bureaucracy, and corruption (Kurt, 2023: Kyle et al., 2018). Beneath its democratic appeal, though, is frequently a divisive language that feeds on animosity, fear, and anger toward those who are viewed as outsiders. The moral limits of public discourse have been altered and hate speech has become more accepted as a result of this rhetoric's increased dissemination through digital media (Wodak, 2015; Farkas & Schou, 2018).

Populism's emotional connection and simplicity make it a powerful communicator. To engage big audiences, populist leaders frequently use symbolic communication and straightforward, passionate language. They portray minorities, immigrants, academics, and political opponents as dangers to the country's moral and cultural fiber while claiming to speak for "real people (Krzyżanowski, 2020; Moffitt, 2016)." This "us versus them" dichotomy reduces complicated socioeconomic issues to moral conflicts, and populist rhetoric serves as a tool for marginalization and stigmatization. The language that results normalizes animosity and offers a moral defense of discriminating against beliefs that might otherwise be considered intolerable.

These dynamics have been exacerbated by the emergence of digital media, which provides populist actors with a decentralized, unmediated platform for direct audience communication. Emotionally charged content spreads quickly thanks to social media networks, online discussion boards, and algorithm-driven environments, which frequently prioritize indignation over thoughtful analysis (Gerbaudo, 2018). Populist politicians like Recep Tayyip Erdoğan, Jair Bolsonaro, Donald Trump, and other right-wing European leaders have adeptly used digital tools to create devoted online groups that reinforce and magnify their divisive views (Freelon et al., 2020; Kurt 2024). Under the cover of "free expression" or "patriotic truth-telling," hate speech is made more visible and socially acceptable by this digital populism.

Furthermore, populist rhetoric's impact extends beyond communication. It reshapes social relations, weakens trust in institutions, and erodes the shared norms that sustain democratic societies. When hate-laden narratives dominate digital discourse, they contribute to polarization, marginalization, and even violence. The stigmatization of vulnerable groups—whether migrants, religious minorities, or political dissidents—undermines human rights and threatens the social cohesion necessary for peaceful coexistence (United Nations, 2019).

Analyzing how political rhetoric impacts societal harmony, how populist communication legitimizes hate speech, and how online platforms both support and undermine these trends becomes critical in this environment. Thus, this study explores populist groups' communication tactics and their wider ramifications for human rights and democratic integrity. In addition to exploring the potential for reestablishing civility and inclusiveness in public discourse, it aims to offer a nuanced understanding of the ways in which digital populism turns language into a vehicle of divide.

The research specifically addresses three interrelated questions:

1. How do populist communication strategies legitimize hate speech?
2. What is the impact of political rhetoric on societal cohesion?
3. How do online platforms enable or restrain the spread of populist hate narratives?

In order to contribute to the ongoing scholarly and policy debates regarding the regulation of hate speech and the preservation of democratic communication in the digital age, this paper will examine these questions through an interdisciplinary lens that combines insights from political science, sociology, psychology, and human rights studies.

2. Conceptual Framework: Populism, Hate Speech, and Digital Communication

According to Mudde and Kaltwasser (2017), populism is a political ideology and communication tactic that portrays society as being split between two opposing factions: the honest people and the dishonest elite. Although populist rhetoric is not intrinsically anti-democratic, it frequently serves as a vehicle for the mobilization of animosity and exclusion. Fundamentally, populism reduces complicated social and political concerns to emotionally charged stories that play on feelings of annoyance, anxiety, and self-identity (Salmela& von Scheve, 2017). Supporters' sense of belonging can be reinforced by this emotional appeal, which also serves to justify animosity toward alleged outsiders.

To understand how hate speech becomes accepted in public life, a good understanding of the communication aspect of populism is required. Populist leaders use performative acts of communication in addition to political programs to create significance. They assert moral superiority and condemn opponents as enemies of the country, religion, or tradition by presenting themselves as the true voice of "the people (Krzyżanowski, 2020)." Moral hierarchies are established by this process of others, which stigmatizes specific social groups as dangers to societal stability and cultural purity. These groups are frequently migrants, ethnic minorities, or political dissidents. Ordinary people are turned into participants in symbolic battles for identity and belonging by the language of fear and rage (Moffitt, 2016).

2.1. Populism and the Logic of Stigmatization

The core of populist discourse is the logic of stigmatization. Stigma serves as a social control mechanism, designating and excluding people who are thought to be inferior or different. By encouraging followers to see themselves as victims of a corrupt system and to focus their resentment on a scapegoated "other," populist rhetoric uses this technique to keep followers emotionally united. In addition to dividing society, this rhetorical device also affects how the general public views authority, morality, and the truth.

According to Kurt (2023), stigma has the power to change how people view the world and rewrite social narratives, which can have an impact on how societies are structured and how people interact with one another. The boundaries of acceptable discourse are gradually shifted by the frequent use of pejorative terms and exclusionary metaphors in political communication, such as "invaders," "traitors," or "fake patriots." Repetition and emotional reinforcement normalize things that were previously deemed offensive. Therefore, populism, frequently in the name of upholding traditional values or defending the country, legitimizes stigmatization by turning prejudice into a moral virtue (Wodak, 2015).

2.2. Digital Platforms and the Amplification of Hate

Populist rhetoric has gained more traction and influence as a result of the digital revolution in political communication. Populist actors have direct, unfiltered access to the audience through social media sites like Facebook, YouTube, and X (formerly Twitter) (Freelon et al., 2020). Digital spaces thrive on virality and engagement, frequently giving preference to emotionally charged, contentious, and sensational messages, in contrast to traditional media, where content is regulated by gatekeeping and editorial standards.

In this procedure, algorithms are essential. Digital platforms produce echo chambers where users are constantly exposed to affirming information by giving preference to content that elicits strong emotional responses. By isolating people inside ideologically homogeneous groups, confirming their prejudices, and limiting exposure to opposing viewpoints, these computational dynamics exacerbate polarization (Farkas & Schou, 2018; Kurt, 2023). Hate speech can spread freely in these virtual communities under the guise of "patriotic defense" or "truth-telling". Such language has become more commonplace, which erodes civic trust and democratic discourse by fostering an intolerance-based culture

(Müller & Schwarz, 2021).

Furthermore, people can express animosity without fear of social repercussions due to the anonymity and speed of online communication, which lowers accountability. Thus, "micro-populism," a pervasive kind of populist attitude that permeates ordinary discussions, memes, and viral narratives in addition to formal politics, thrives in the digital sphere (Gerbaudo, 2018). This change makes hate speech a cultural phenomenon as well as a political one, obfuscating the distinction between social radicalization and political mobilization.

2.3. Human Rights Perspective

Human rights are seriously threatened by the confluence of stigmatization, populism, and digital communication. International human rights law's core tenets of equality, nondiscrimination, and human dignity are all violated by the spread of hate speech (United Nations, 2019; European Commission, 2023). However, defending the right to free speech frequently makes attempts to control offensive speech more difficult. Populist actors usually take advantage of this contradiction by attacking vulnerable communities while posing as advocates of "free speech" (Farkas & Schou, 2018).

Addressing digital populism from a human rights standpoint necessitates striking a compromise between preventing harm and preserving the right to free speech. This entails creating thorough frameworks that comprise:

- Legal measures to hold individuals and platforms accountable for hate speech and incitement (European Commission, 2023).
- Educational initiatives to promote media literacy and critical thinking (UNESCO, 2022);
- Civic engagement programs that encourage dialogue and empathy among diverse groups (Norris & Inglehart, 2019; Kurt, 2023).

Moreover, it is impossible to separate the battle against hate speech from larger social and economic injustices. People who feel ignored by political institutions or shut out of economic opportunities frequently connect with populist narratives. Therefore, tackling the root causes of populists' exploitation—poverty, unemployment, social exclusion, and a decline in institutional trust—is essential to advancing human rights in the digital era.

2.4. Toward an Interdisciplinary Understanding

An interdisciplinary perspective is necessary to comprehend how populism, hate speech, and digital communication are all convergent. Human rights studies give normative frameworks to counteract the negative consequences of populist narratives (Mudde & Kaltwasser, 2017; Moffitt, 2016), sociology and psychology unveil the social and emotional mechanisms that make them attractive, and political science sheds light on the tactics and structure of populist groups.

This research aims to shed light on how digital populism alters communication ethics and calls into question the fundamental tenets of democratic coexistence by including these viewpoints. Building on this conceptual framework, the ensuing sections examine the actual expressions of populist hate narratives in several political contexts and suggest inclusive, human rights-based tactics to mitigate their effects.

3. Populist Communication and the Normalization of Hate Speech

Emotionally charged, oversimplified, and moralized narratives that split the political landscape into two opposing camps—the good people and the corrupt elite—are the foundation of populist communication (Mudde & Kaltwasser, 2017). This dichotomous reasoning, however, does not end there; it also applies to people who are depicted as "enemies of the people" or "outsiders." By portraying exclusion and discrimination as moral obligations, brave deeds, or patriotic duties, populist politicians use communication tactics to justify hate speech. Hate speech is normalized—even celebrated—in the public domain by means of these methods (Kurt, 2023; Moffitt, 2016).

Scapegoating, or assigning blame for complicated societal issues on a particular group or society, is a key component of populist discourse (Goffman, 2009). Moral deterioration, economic suffering, or national instability are often attributed to immigrants, religious minorities, LGBTQ+ people, and political dissidents (Wodak, 2015; Norris & Inglehart, 2019).

This rhetorical technique shifts resentment toward weaker targets by transforming structural or policy shortcomings into moral disputes. For instance, building the wall between the United States and Mexico during Donald Trump's administration represented the protection of national identity against a purported "invasion" and went beyond simple policy (Norris & Inglehart, 2019). Under Jair Bolsonaro, similar dynamics developed in Brazil, where journalists, environmental campaigners, and minorities were portrayed as dangers to national sovereignty or traditional values (Gerbaudo, 2018). Recep Tayyip Erdoğan frequently portrays political opponents, Kurds, and secularists in Turkey as internal adversaries who plot to undermine the country's unity (Mulvey, 2025). To rally people, far-right populist movements throughout Europe make use of anxieties about immigration and Islamization.

Scapegoating functions as a short communication cut in each of these situations, simplifying reality, directing public ire, and instilling a phony sense of unity among supporters. By converting prejudice into a shared identity, it enables people to feel that their animosity is ethically justifiable.

Populist discourse is intensely sentimental. Populist leaders appeal to feelings of fear, rage, and resentment rather than using reasoned reasoning (Salmela & von Scheve, 2017). The impression that the country is under siege and that only strong leadership can restore order and purity is one of the carefully constructed feelings that are meant to arouse feelings of crisis and urgency.

Hate speech becomes "legitimate" inside this emotional economy because it is viewed as an act of truth-telling or resistance against censorship and political correctness (Wodak, 2015; Kurt, 2023). When populist leaders make fun of minorities or disparage opponents, their supporters see these acts as displays of bravery and honesty rather than verbal abuse.

This emotional dynamic is amplified by digital media. Short films, slogans, and images distill complicated reality into emotionally charged, readily shared symbols. This content is then rewarded by algorithms, which promote postings that elicit strong reactions, particularly outrage (Vosoughi et al., 2018). Consequently, anger is turned into amusement via internet communication, which instills animosity in regular political conversation.

The way populist speech is received and disseminated has changed due to online media. Echo chambers, which are self-reinforcing environments where users only come across content that supports their own opinions, are made possible by social media. There is little opposition to hate speech in these virtual communities (Sunstein,2018).

By organizing internet campaigns that combine false information, conspiracy theories, and emotionally charged material, populist players exploit this setting. For example, concerted disinformation campaigns depicted Muslims, refugees, or left-wing activists as existential threats during significant election campaigns in the U.S., Brazil, and portions of Europe.

Platforms like Facebook, YouTube, and TikTok favor sensational and polarizing content due to their algorithmic design (Ribeiro et al., 2021). This bias in technology makes hate speech a lucrative and viral commodity by expanding the reach of populist rhetoric. As a result, what starts out as fringe rhetoric gains traction and is accepted by both political elites and the design of digital communication.

Social standards are progressively altered by populist discourse as it permeates daily media and political discussions. Racist jokes, disparaging comments, or overt xenophobia—things that were formerly deemed unacceptable—become "normal" or "authentic". Hate speech is justified as free speech, and political correctness is reframed as censorship (Urbinati, 2019).

It is frequently difficult for institutions like the media, academics, and the court to react in an efficient manner. Occasionally, populist administrations deliberately weaken these establishments by designating them as "enemies of the people." Further radicalization is made possible by this delegitimization, which also erodes democratic protection.

In Turkey, for instance, critical journalists and academics have been characterized as conspiracies or "foreign agents." To suppress opposition, populist regimes in Hungary and Poland have reorganized media outlets and employed nationalist rhetoric. Such events demonstrate how, once accepted, hate speech undermines accountability and freedom while posing as a defense.

Human rights are significantly impacted by the normalization of hate speech. The legitimization of disparaging speech frequently results in physical violence, harassment, and discrimination (Gelber, 2021). Hate speech serves as a prelude to hate crime, which is a process by which dehumanization comes before exclusion or violence.

Populist regimes, by stigmatizing disadvantaged populations, create an environment where human rights breaches become socially and politically acceptable. The fundamental idea that all people are created equal in terms of their rights and dignity is compromised by this blurring of moral and legal lines.

Therefore, the battle against digital populism must address the ideological and structural underpinnings of hatred, such as institutional mistrust, economic inequality, and the exploitation of cultural fear, in addition to regulating hate online.

4. The Impact of Political Rhetoric on Social Cohesion and Democratic Integrity

Politics' language creates reality rather than just describing it(Fairclough, 2013). Political speech has become a key component in determining how societies' collective conceptions, identities, and connections are shaped, especially in populist movements. The moral foundation that underpins democratic cooperation is undermined when political discourse increasingly turns to divisive and exclusive narratives (Mudde & Kaltwasser, 2017). It will be examined here how populist language erodes civic trust, threatens social cohesiveness, and turns democratic discourse into a battleground of animosity and hatred.

Political communication should encourage discussion, compromise, and deliberation in robust democracies (Habermas, 1996). But populist discourse substitutes hostility for this pluralistic paradigm. Politics is delegitimized and disagreement is turned into animosity by portraying it as a conflict between "the pure people" and "the corrupt elite."

This change is an example of symbolic violence, which is the process by which representations and words cause moral injury (Bourdieu, 1991). Political leaders alter the social boundaries of inclusion and exclusion when they refer to journalists as "enemies of the people," migrants as "invaders," or critics as "traitors." Such verbal anger justifies physical violence against marginalized people in addition to verbal animosity (Gidron & Bonikowski, 2013).

This change is best illustrated by Donald Trump's designation of the media as "fake news" and "the enemy of the people" (Norris & Inglehart, 2019). Similar to this, populist groups in Europe and Latin America use derogatory language to stigmatize minorities, refugees, and LGBTQ+ people, which feeds into cycles of prejudice and division. Political discourse in Turkey has progressively reduced the scope of public discourse by portraying civil society activists and dissenters as foreign conspiracies or "terror sympathizers."

Trust is essential to social cohesiveness, both between individuals and between institutions (Putnam, 2000). Both are undermined by populist speech. Populist politicians undermine accountability and collective governance by consistently portraying institutions like the media, courts, and universities as biased or corrupt.

When trust is damaged, people withdraw into ideological camps characterized by mistrust and terror rather than participating in cooperative public life. The end effect is a fractured society in which people view one another as enemies rather than collaborators in a common democratic endeavor.

For example, studies conducted in Brazil and the United States show that party affiliation has progressively supplanted civic identity, creating a strong feeling of moral and cultural divisiveness (Norris & Inglehart, 2019). Populist regimes have strengthened majoritarian nationalism at the price of pluralism in Turkey and other Central European countries, encouraging conformity and suppressing dissent (Mulvey, 2025). The inclusive basis of democratic solidarity is weakened by this dynamic.

Though in a warped form, populist discourse takes use of identity politics (Brubaker, 2017). Populist actors use identification as a means of divide, whereas progressive identity groups aim for equality and acknowledgment. They create "imagined communities" centered on cultural, religious, or ethnonationalist lines, claiming that outsiders or liberal elites are attacking these identities.

By generating fragmented micro-publics—online venues where individuals consume content specific to their identification group—digital platforms exacerbate this dynamic. Algorithms strengthen the gap between "us" and "them," rather than encouraging communication across divides (Tucker et al., 2018).

As a result, what was formerly thought of as a forum for logical discussion turns into a network of echo chambers controlled by ideology and emotion. In addition to undermining mutual understanding, this digital fragmentation makes reaching a democratic consensus more difficult (Krzyżanowski, 2020).

When inflammatory discourse is repeated, psychological desensitization results. People become used to verbal abuse and accept it as a common occurrence in public conversation. As cynicism and animosity take the place of empathy and compassion, this desensitization adds to the moral decline of civilizations (Haidt, 2012).

Furthermore, historical awareness and communal memory are altered by ongoing exposure to hate speech. In order to exalt the country's past and demonize minorities or foreign powers, populist politicians frequently alter history. By doing this, they create emotionally charged narratives that defend repression and discrimination under the guise of national rejuvenation or cultural purity.

In addition to fostering bias, this distortion of historical narratives obstructs intergenerational understanding. Younger generations may adopt intolerance as a valid form of civic engagement if they are raised in surroundings where hate speech is prevalent (Hertzoff, 2025).

Populist discourse has the cumulative impact of democratic backsliding, which is the slow deterioration of civic liberties, pluralistic standards, and checks and balances (Levitsky & Ziblatt, 2018). Under the pretense of defending the "will of the people," political leaders can consolidate power and stifle opposition by inciting hatred and terror.

Viktor Orbán's talk of "illiberal democracy" in Hungary serves as an example of how populist rhetoric in Hungary justifies authoritarian policies (Pappas, 2019). Erdoğan has frequently used religious morality and national unity in order to bolster his authority over the press, the court, and civil society in Turkey (Kurt, 2023). Trump's "Make America Great Again" campaign created an atmosphere where electoral legitimacy and institutional norms were publicly challenged on the other side of the Atlantic (Norris & Inglehart, 2019).

These instances show how populist discourse serves as a sign of democratic deterioration as much as its cause. It erodes the civic underpinnings necessary for democratic cooperation by normalizing animosity and weakening respect for one another.

A multifaceted approach is necessary to address the effects of populist discourse. Although vital, legal prohibition of hate speech is insufficient (UNESCO, 2022). Societies must also make investments in ethical communication, or public debate based on respect, empathy, and critical thinking.

Rebuilding social cohesion and trust requires civic education, interfaith and intercultural discussions, and responsible media literacy initiatives (Hertzoff, 2025). Promoting forums for discussion that cut across ideological divides ought to be the responsibility of academic institutions and civil society groups.

In the end, combating hate speech is simply one way to counteract populist division; another is to foster an inclusive narrative of our common humanity, one that views variety as a strength rather than a danger.

5. Digital Platforms and the Dynamics of Online Hate

The production, dissemination, and consumption of political ideas have all been completely transformed by the digital revolution in communication. Online platforms have made it easier for people to participate and access information, but they have also made it easier for hate speech and populist propaganda to spread (Castano-Pulgarín et al., 2021). Digital spaces have become effective instruments for political manipulation due to attention-based economics, algorithms that emphasize participation, and lax regulatory oversight (Gillespie, 2018). This part looks at how internet platforms facilitate, legitimize, and occasionally limit the spread of hateful populist narratives.

Social media sites like YouTube, Facebook, TikTok, and X (previously Twitter) are not impartial middlemen. Because strong emotional responses—such as wrath, outrage, or fear—drive engagement, its design naturally favors material that evokes these kinds of feelings (Brady et al., 2017; Vaidhyanathan, 2018).

This digital logic has been mastered by populist movements and politicians. They take advantage of the computational architecture that favors virality over truth by producing divisive, emotionally charged statements.

Hate speech and conspiracy theories spread more quickly than accurate, balanced information because of this architecture of amplification, which turns contentious language into entertainment (Marwick & Lewis, 2017). Such content becomes more visible the more people respond, share, or comment on it. As a result, digital media influences the environment in which political discourse takes place in addition to spreading populist themes.

In the attention economy in which digital platforms function, visibility is equivalent to value (Wu, 2016). In order to optimize online time and turn interaction into ad income, algorithms are constantly learning from user behavior (Zuboff, 2019).

By creating dramatic, conflict-driven material that captivates viewers, populist leaders take advantage of this economy.

The populist goal unintentionally coincides with this capitalist logic. Conspiracy theories, hate speech, and moral panic all lead to high engagement, which makes platforms money. Thus, the politics of hatred and the economics of attention are intertwined.

Populist actors gain legitimacy by their online presence; the more often they show up in feeds, the more people view them as genuine representatives of "the people" (Mudde & Kaltwasser, 2017). They frequently silence moderate views because to their dominance in the digital public realm, undermining pluralism and turning public debate into a marketplace of indignation.

Algorithms reinforce established ideas by filtering and personalizing material based on user preferences (Tambini, 2021). This eventually results in algorithmic polarization, or echo chambers where individuals are only exposed to viewpoints that align with their own.

Populist rhetoric flourishes in these echo chambers. Misinformation is allowed to flourish unchecked because hateful or conspiratorial comments are rarely contested (Guess et al., 2019). Because people mistake the uniformity of their online networks for broad consensus, this digital isolation encourages radicalism.

Recommendation algorithms on websites like YouTube and TikTok frequently steer people away from popular political content and toward more extremist or hateful content (Ribeiro et al., 2021). Although the algorithm does not "intentionally" radicalize, people are inexorably drawn to emotionally charged content since it is optimized for interaction.

As a result, the digital realm splits into ideological hotspots that erode social cohesiveness and increase intergroup mistrust.

Governments and social media corporations have started enacting laws to stop hate speech in recent years (UNESCO, 2022). Regulating hate speech online is still difficult, though.

First, enforcement is made more difficult by the worldwide reach of digital communication. In one legal system, what is considered hate speech could be protected speech in another (United Nations, 2019). Second, overly stringent regulations run the danger of violating the fundamental democratic value of freedom of expression (Tambini, 2021). Third, platforms' content moderation procedures are sometimes opaque, which raises questions about censorship, prejudice, and political manipulation (Gorwa et al., 2020).

The Digital Services Act (DSA) of the European Union and the NetzDG of Germany are significant initiatives to hold digital corporations accountable by mandating the prompt removal of unlawful information and more algorithmic transparency. However, rather of being preventive, these initiatives continue to be reactive.

Eliminating dangerous content is only one aspect of true digital control. It ought to encourage civic education, algorithmic accountability, and ethical design, all of which enable people to interact critically with information.

Digital platforms represent a paradox: they allow people to express themselves, but they also make them vulnerable to division and manipulation. Although they give voice to underrepresented communities, they also act as loudspeakers for populist demagogues who take use of these same liberties to promote violence and hatred (Urbinati, 2019; Kurt, 2023).

This contradiction illustrates a more fundamental moral conundrum facing contemporary democracies: how to maintain free speech without allowing it to be used as a weapon.

A democratically valid regulatory framework that protects human dignity without compromising intellectual freedom is necessary to resolve this conflict.

Three fundamental ideas must be incorporated into every effective solution:

1. Transparency – Platforms should disclose how algorithms prioritize and moderate content.

2. Accountability – Companies must be held responsible for the social impact of their technologies.

3. Empowerment – Users must be equipped with digital literacy to identify manipulation and resist polarization.

Societies need to develop a new kind of digital citizenship to stop the propagation of hate online (Binny et al., 2019; Mihailidis, 2018). This means encouraging consumers to think critically, be empathetic, and act ethically.

Building resilience against populist narratives requires educational initiatives that teach youth how to avoid false information, value diversity, and have productive conversations.

Furthermore, cooperation between governments, academic institutions, civil society organizations, and IT firms is essential. The problem is cultural as well as technological. Any legislation that lacks a common moral commitment to human dignity will remain shallow.

The most effective remedy for the poisonous dynamics of digital populism is a culture of responsible communication that is based on respect for one another and civic consciousness.

6. Regulation, Responsibility, and the Protection of Human Dignity

One of the most difficult problems facing contemporary democracies is the control of hate speech online. The protection of fundamental rights, most notably freedom of expression, must coexist with the need to keep people and groups safe. The challenge for policymakers, academics, and civil societies is to create frameworks that are both successful and politically acceptable as populist leaders and movements use internet communication as a weapon to spread hatred (Moffitt, 2016).

This part integrates knowledge from political science, communication studies, law, and human rights ethics to present a normative and useful framework for dealing with hate speech in the digital age.

The dichotomy between unrestricted freedom and official censorship frequently ensnares attempts to control hate speech. Neither extreme is sufficient. While unregulated areas allow hatred to spread, overly restrictive ones run the danger of stifling criticism and undermining democratic liberties.

Therefore, the goal of a democratic framework must be to foster an environment that encourages discussion, plurality, and respect rather than to stifle expression. Legitimate political speech should not be restricted by regulations that shield people from dehumanization.

This calls for a change in approach from reactive deletion to proactive prevention, with a focus on ethical responsibility, openness, and user empowerment (Gillespie, 2018).

A sustainable approach to combating digital hate must rest on three interrelated principles:

1. Human Dignity as a Core Value

The protection of human dignity, the cornerstone of all human rights, must be given top priority in regulations (United Nations, 2019). Since hate speech aims to exclude and humiliate people by definition, protecting dignity entails limiting such communication while yet allowing for free discussion (UNESCO, 2022).

2. Proportionality and Accountability

Prohibitions of hate speech must be reasonable, responsive to the situation, and democratically supervised. Both corporate platforms and governments must answer for the impact of their actions on civil freedoms and public discourse (Gorwa, 2019).

3. Transparency and Due Process

Users ought to have the right to know how algorithms filter information, why content is deleted, and how choices about moderation are made. Building trust and ensuring that regulations serve the public interest rather than corporate or political goals are two benefits of transparent government (European Commission, 2023).

Governments are essential in defining the moral and legal guidelines for online communication. The General Data Protection Regulation (GDPR) and the Digital Services Act (DSA) of the European Union offer useful examples of how to strike a balance between user protection and personal liberty (European Commission, 2023).

But in a digital world with no borders, national policies are not enough. Hate speech frequently crosses national boundaries, necessitating international cooperation. Organizations like the United Nations and the Council of Europe can promote uniform norms that safeguard free speech while outlawing incitement to violence and hatred.

However, it is important to avoid political abuse of anti-hate laws, which might be used for authoritarian purposes while masquerading as moral defense (Norris & Inglehart, 2019).

Digital platforms serve as private regulators of public conversation, whether they are new networks or multinational companies like Meta and Google (Gillespie, 2018). Their decision-making is frequently unaccountable, despite their enormous potential to influence public opinion.

In order to establish a more democratic digital environment, platforms need to:

- Disclose algorithmic criteria for content ranking, recommendation, and moderation.
- Establish independent oversight boards that include scholars, ethicists, and civil society representatives.
- Invest in counter-speech initiatives that promote diversity, empathy, and factual accuracy.
- Develop ethical AI frameworks that minimize bias and reduce the amplification of divisive or hateful content.

In a world where digital design shapes political reality, corporate responsibility and algorithmic transparency are not just technological concerns (Elliott et al., 2021); they are moral requirements.

A knowledgeable and compassionate populace, rather than just regulations, is the most resilient defense against hate speech on the internet. It is imperative that academics, cultural institutions, and civil society groups collaborate to create digital literacy initiatives by identifying false information and deceptive populist language. People can learn how to have civil conversations across ideological differences and effectively and ethically report hate speech.

Thus, resilience is built on education. Both authoritarian populism and the allure of hatred may be resisted by a culture that values empathy, critical thinking, and respect for one another.

Laws and algorithms alone will not be sufficient to defeat digital hate speech; a shift in communication culture is needed. The ethics of dialogue—respect for diversity, the pursuit of truth, and the preservation of our common humanity—must be internalized by people, media organizations, and political actors alike.

Where fear and frustration are prevalent, populist discourse flourishes. Societies may counteract the emotional attraction of divided narratives by reestablishing faith in democratic institutions and highlighting human unity.

The establishment of forums where debate does not turn into dehumanization is essential to an ethical public sphere, both online and offline.

Moreover, finding a balance between freedom and responsibility is necessary to promote democracy. One of the most important tests for modern democracy is the policing of hate speech on the internet (Rodríguez-Peralet al., 2025). Creating a communication climate where all voices are heard without fear or hatred is the task, not stifling them.

Therefore, a democratic framework for combating hate speech and online populism must be incorporated:

- Legal clarity, to define and sanction harmful conduct.
- Technological responsibility, to ensure algorithms serve human dignity.
- Civic empowerment, to build awareness, dialogue, and resilience.

Societies can only progress toward a digital future that preserves freedom, protects human rights, and reestablishes the moral underpinnings of public life by bringing these three dimensions—law, technology, and education—together.

Fighting online hatred is ultimately a moral commitment to the notion that how we choose to talk to and about one another determines the fate of democracy itself, not just a regulatory endeavor.

7. Conclusion

This study has investigated how internet platforms influence the spread of populist hate narratives, how political rhetoric affects social cohesiveness, and how populist communication tactics legitimate hate speech. It has been demonstrated through an interdisciplinary approach that integrates political science, sociology, psychology, and human rights studies that the emergence of digital populism is both a sign of and a cause of the deterioration of democracy.

Populist movements and leaders create a moral division between "the corrupt elite" and "the pure people." By doing this, they turn justifiable complaints into sentimental stories that arouse rage, fear, and bitterness. By taking use of digital channels, they use affective communication—such as memes, hashtags, slogans, and emotionally charged videos—to magnify these feelings, circumventing logical thought and promoting animosity based on identity.

There are two main ways in which this rhetorical dynamic validates hate speech. First, it makes prejudice socially acceptable by normalizing stigmatization, which portrays migrants, minorities, or dissenters as dangers to national identity. Second, by using freedom of speech as a weapon, populists assert their moral superiority and intentionally undermine pluralism while presenting censorship resistance as a democratic defense.

Social cohesiveness is weakened as a result, with tribalized speech and digital echo chambers replacing common narratives of empathy and belonging. In addition to undermining interpersonal trust, polarization and stigmatization foster authoritarian inclinations.

Digital platforms have a conflicting role in this change. While their algorithms encourage controversial material and favor emotional extremes, they also democratize access to knowledge and facilitate civic engagement. Thus, the exact populist processes that propagate hatred are maintained by the social media architecture, which is based on virality and engagement metrics.

A multifaceted democratic framework that strikes a balance between freedom and responsibility is needed to address these issues. Beyond harsh penalties, regulation must incorporate moral principles of openness, responsibility, and respect for human dignity. Citizens must be prepared by education to critically and sympathetically traverse information environments. As a moral counterbalance, civil society must establish forums for discussion and inclusiveness.

Fighting the echoes of hatred in the digital era is ultimately a cultural and ethical task rather than just a technological or legal one. Democracies must reassert that human rights continue to be the cornerstone of all acceptable communication, that variety unites rather than divides, and that the right to free speech entails a duty to respect. Democracy's future rests not just on who can talk loudly but also on whether or not we can still listen to each other with respect and understanding.

References

Binny, M., Ritam, D., Pawan, G. and Animesh, M. (2019). 'Spread of Hate Speech in Online Social Media' *Proceedings of the 10th ACM Conference on Web Science* 173, https://doi.org/10.48550/arxiv.1812.01693 (accessed 28 sep. 2025).

Bourdieu, P. (1991). *Language and symbolic power* (J. B. Thompson, Ed.; G. Raymond & M. Adamson, Trans.). Harvard University Press.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences,* 114(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brubaker, R. (2017). Between nationalism and civilizationism: The European populist moment in comparative perspective. *Ethnic and racial studies* 40(8): 1191–1226. https://doi.org/10.1080/01419870.2017.1294700.

Castano-Pulgarín, S. A., Suárez-Betancur, N., Tilano Vega, L. M., & Herrera López, H. M. (2021). Internet, social media and online hate speech: Systematic review. *Aggression and Violent Behavior*, 58, 101608. https://doi.org/10.1016/j.avb.2021.101608

Elliott, K., Price, R., Shaw, P. et al. (2021). Towards an Equitable Digital Society: Artificial Intelligence (AI) and Corporate Digital Responsibility (CDR). *Soc* 58, 179–188 https://doi.org/10.1007/s12115-021-00594-8

European Commission (2023). '*No Place for Hate in Europe. Commission and High Representative Launch Call to Action to Unite Against All Forms of Hatred'* https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6329 (accessed 28 May 2024).

Fairclough, N. (2013). *Critical discourse analysis: The critical study of language* (2nd ed.). Routledge.

Farkas, J., & Schou, J. (2018). *Post-truth, fake news and democracy: Mapping the politics of falsehood*. Routledge.

Freelon, D., Marwick, A., & Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science*, 369(6508), 1197–1201. https://doi.org/10.1126/science.abb2428

Gelber, K. (2021) 'Differentiating Hate Speech: A Systemic Discrimination Approach' 24:4 *Critical Review of International Social and Political Philosophy* 393–414. https://doi.org/10.1080/13698230.2019.1576006 393–414.

Gerbaudo, P. (2018). *The digital party: Political organisation and online democracy*. Pluto Press.

Gidron, N., & Bonikowski, B. (2013). V*arieties of Populism: Literature Review and Research Agenda* (Weatherhead Working Paper Series, No. 13-0004).

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gorwa, R. (2019). *What is platform governance? Information, Communication & Society*, 22(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Guess, A. M., Nagler, J., & Tucker, J. A. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. MIT Press.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage Books.

Hertzoff, A. (2025). Anger and Modern Politics. In: *Anger in Politics*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-94713-1_7

Krzyżanowski, M. (2020). Discursive shifts in ethno-nationalist politics: On the theory and practice of populist discourse. *Journal of Language and Politics*, 19(4), 523–546. https://doi.org/10.1075/jlp.19076.krz

Kurt İ. (2024). *A long walk to Human Rights. Journal of Human Rights and Refugee Studies*, ISSN 2944-2656, 11-29.

Kurt, I. (2023). *Populism, stigmatization, and the moral economy of hate: Digital populism and its human rights implications. Populism Reloaded?* Revistia Publishing

And Research, Proceedings Book ISBN 978-1-915312-08-2

Kyle, J., Gultchin, L., and Gultchin, L., (2018). *Populism in Power Around the World*. Available at SSRN: https://ssrn.com/abstract=3283962 or http://dx.doi.org/10.2139/ssrn.3283962

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. Crown Publishing Group.

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online.* Data & Society Research Institute.

Mihailidis, P. (2018). *Civic media literacies: Re-imagining human connection in an age of digital abundance*. Routledge.

Moffitt, B. (2016). *The global rise of populism: Performance, political style, and representation*. Stanford University Press.

Mudde, C., & Kaltwasser, C. R. (2017). *Populism: A very short introductio*n. Oxford University Press.

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4), 2131–2167. https://doi.org/10.1093/jeea/jvaa045

Mulvey, J. (2025). *How US intelligence and anti-communism have shaped Turkish Politics to this day*, CeSPI Intern

Norris, P., & Inglehart, R. (2019). *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press.

Pappas, T. S. (2019). *Populism and liberal democracy: A comparative and theoretical analysis*. Oxford University Press. https://doi.org/10.1093/oso/9780198837886.001.0001

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. Simon & Schuster.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira Jr., W. (2021). Auditing radicalization pathways on YouTube. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3442188.3445919

Rodríguez-Peral, E., Gómez Franco, T., & Rodríguez-Peral Bustos, D. (2025). Propagation of Hate Speech on Social Network X: Trends and Approaches. *Social Inclusion*, 13, Article 9317. https://doi.org/10.17645/si.9317

Salmela, M., & von Scheve, C. (2017). Emotional roots of right-wing political populism. *Social Science Information*, 56(4), 567–595. https://doi.org/10.1177/0539018417734419

Sunstein, C. R. (2018). *Republic: Divided democracy in the age of social media.*

Princeton University Press.

Tambini, D. (2021). *Media freedom and the regulation of online platforms*. Council of Europe.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political Science Quarterly*, 133(3), 707–749.

UNESCO (2022). *Addressing hate speech through education*. UNESCO Publishing.

United Nations (UN) (2019). *United Nations Strategy and Plan of Action on Hate Speech*. https://www.un.org/en/hate-speech

United Nations (UN) (2019). *Report on combating hate speech.* Office of the UN High Commissioner for Human Rights.

Urbinati, N. (2019). *Me the people: How populism transforms democracy*. Harvard University Press.

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wodak, R. (2015). *The politics of fear: What right-wing populist discourses mean*. Sage Publications.

Wu, T. (2016). *The attention merchants: The epic scramble to get inside our heads*. Knopf.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Digital Hate to Real World Harm: Online Platforms, Populist Leader Rhetoric and The Escalation of Hate Crime

Sahil Lohia, Jamia Hamdard University

1. Introduction

Across existing democracies, hate speech has transformed from a fringe communication phenomenon into a digitally driven real world violence. The growth of digital platforms has fundamentally changed how hostility is articulated, circulated and acted upon. In comparison, traditional forms of hate required physical proximity, organisational coordination and influential intermediatory. Digital infrastructure enables populist leaders and micro populist influencers to broadcast hate to millions of global citizens in seconds. These speeches, when algorithmically amplified, acquire moral force: it normalise hostility, legitimises organised aggression and allows violence to be reframed as a moral civic duty rather than criminal conduct (Bandura, 1999; Benesch, 2014). The escalation in populist rhetoric is not accidental. It is a by-product of digital platform structure that rewards sensational, emotionally charged content. (Sunstein, 2017; Gillespie, 2018). This populist rhetoric is now disassembled into a 30 second video clips that compress the erstwhile newsletters into a cluster of emotional triggers. Populist leaders have played a key role in this shift. They frame their messages around a binary moral world: a pure people and a corrupt enemy (Mudde, 2004). In this discursive world, minorities, immigrants, religious groups, tribes, political opponents are just competitors but existential threats. These narratives create "permission structures", Discursive atmosphere in which ordinary citizens begin to treat violence as the only rational response to imaginary danger (Bandura, 1999). This paper examines how hate emerges online, how platforms magnify it and how rhetoric translates into a real-world hate crimes. It further argues that digital platforms do not distribute political content evenly, However the algorithmic incentive system rewards it in the form of communicative violence: the more provocative or outrageous a speech is, the more its distribution is amplified. India offers the clearest global cases as its digital ecology is uniquely powerful, as it merged a large population with cheap internet, high rate of smartphone penetration, decentralised group encrypted messaging dominated by WhatsApp (in these groups speeches are already validated by kinship) (Chadha & Guha, 2016). India demonstrates the most intense connection between rhetorical violence and real-world harm via riots, lynching, communal programs and vigilant attacks. The same transformation of credible threats rarely occurs in open high-friction digital spaces like X. Populist figures from The Unites states of America, Brazil and Hungary engage in xenophobic or demonising rhetoric that has been extensively documented (Berman, 2019; Wodak, 2015; Pappas, 2019). This section, therefore, frames digital hate not as a sequence of isolated hateful messages but as a socio technical system involving three components: populist rhetoric, platform amplification and real world mobilisation. These elements create a cyclical model in which extremism once seeded is reproduced and intensifies through feedback loops of spectacle and collective emotion (Ribeiro et al., 2020).

2. Literature and Conceptual Background

The study of hate speech and its relationship with violence sits at the intersection of political communication, criminology, platform governance and digital sociology. Each field approach this phenomenon through a different lenses, but converges on a common point that hate speech is not merely descriptive; it is performative. Hate speech reconstitutes social boundaries, creates permissive moral environments and reframes the ontological status of its targets. In this sense it functions not as a linguistic aggression but as a behavioral script , preparing people to reinterpret aggression as a social necessity (Bandura, 1999; Benesch, 2014).

Hate speech as a Precursor

Criminological models treat hate speech as a precursor mechanism. Through insults, conspiracy narratives, dehumanising labels isolates minority groups from the moral community and gradually lowers inhibitions to harm. Bandura's theory of moral disengagement is foundational here: individuals do not commit violence because they lack ethical framework, rather they commit violence when rhetorical cues allow them to label themselves as protectors and their victims as parasite, traitorous or parasites(Bandura, 1999). These victims are targeted as existential threats rather than citizens, violence recasted as moral duty rather than criminality. Digital circulation intensifies this transformation through repetition, virality and memeification. This process mimics psychological grooming. The first exposure is shocking and tenth is banal. Benesch (2014) terms dangerous speech as language thtat increases the likelihood of violence even without issuing expilict instructions. This distinction is important because digital hate authorises rarely commands. The difference between they are destroying us and go and kill them is legally vast but functionally narrow in this ecosystem.

Populist Rhetoric and "we the People"

Populism operates as rhetorical engine that makes a mythical unity the people as opposed to a designated enemy. This ideological core claim is moral rather than policy driven. In Europe, Viktor Orban repeatedly framed migrants as cultural poison threatening to contaminate Hungarian civilization (Wodak, 2015). Populist leaders rarely demand violence, instead they narrate hostility as a patriotic responsibility. Through this displacement, aggression is recoded from emotion to duty. This rhetorical shift mirrors the moral disengagement processes where aggressors become protectors, minorities become invaders, and violence becomes defence. (Bandura,1999)

2.3 Digitally Networked Hate Patterns.

Digital infrastructure amplifies these rhetorical patterns by converting them into economic incentives. The networked mobilisation demonstrates that digital platforms catalyses mobilisation without traditional organisational hierarchies. Digital platforms also provides an attention economy through which a digital platform rewards emotional extremity because outrage produces comments,shares and time on screen, these behaviors maximises profits (Zuboff, 2019).

Instagram reels and you tube shorts is used to compress complex political discourse into emotionally charged fragements. A speech on citizenship reservation policy or freedom of speech is remixed in to a weaponised soundbite accompanied by dramatic music, nationalist iconography and motivational text overlays. Whatsapp multiplies this escalation through private group networks “digital villages”. Unlike X or youtube, where content is publicly contested. Whatsapp messeges are circulated into extended family,caste group, alumni batches, religious communities (Chakravartty & Roy, 2017; Chadha & Guha, 2016). In this encrypted ecosystem speech is presumed more credible because it arrives through social proximity rather than algorithmic recommendation

Comparative Context: Populists Who Authorise Hate

This section briefly examines international precedents that illustrate how populist leaders authorises violence without commanding it

Vikar Orban (Hungary)

Hungary populist leader anti-immigrant campaigns reframed refugees as existential threats rather than humanitarian subjects. His speech portrayed migrants as toxic infiltrators intended to dismantle European civilisation (Wodak, 2015). State television then reproduced those narratives through sustained informational campaigns resulting in hate crimes against refugees increased in parallel: street act justifies as patriotic act. Later on, Orban defended restriction on civil liberties as necessary to protect Europe (Berman, 2019).

Jair Bolsonaro (Brazil)

Bolsonaro speeches targeted gender minorities, tribes and afro Brazilians communities. Brazilian leader went further to describe indigenous Brazilians as animals. Jair’s rhetoric celebrated cruelty as a masculine courage and transformed social hostility into performance of patriotism. The vigilantes did not quote any ideology, but they quoted him during his tenure activists, land defenders and tribal leaders were murdered at unprecedented rates (Human Rights Watch,2019).

Donald Trump (United States)

Donald trump's anti-immigrant rhetoric escalated from acquisition of criminality to claim soft treason and invasion of American identity. Trump never issued explicit instructions to attack, but he simply questioned democracy, legitimacy and framed resistance as a patriotic duty resulted in on the capital riots in with thousands of supporters physically occupied congress. (Pappas, 2019).

Across these examples, rhetoric did not command violence, but it licences.

India as a Global Case of Digital Hate Escalation

India presents a world's most complex ecology of digitally mediated hate. The origin of hate speech is decentralised: speeches here do flows from regional strongman, caste entrepreneurs' religious influencers. These figures command audiences built through moral charisma symbolic purity not through state machinery. In this digital arena a social influencer with 300,000 digital followers may mobilise as many audiences as a cabinet minister in another country. This algorithmic infrastructure enables WhatsApp forwarding loops, telegrams channels and Instagram reels functions as amplification circuits. Messages pass through broadcasting channels, WhatsApp group and even stories. This information enter communities as assumptions not as arguments (Chakravartty & Roy, 2017).

Memory of Permission: Rajiv Gandhi and the Sikh Riots (1984)

The most enduring textbook case of permission rhetoric in India is aftermath of former Primer minister's Indira Gandhi's assassination in 1984. While addressing a crowd in New Delhi after Sikh riots in Delhi, then prime minister Rajiv Gandhi stated:

"When a big tree falls, the earth shakes."

The sentence in his speech did not condemn it, nor instruct violence. Instead, it offers moral grounding for the Sikh's slaughter. This shift in interpretive horizon continues to echo through Indian political psychology: when elected leaders characterise collective violence as a logical response, vigilantism becomes a form of civic housekeeping. (Staff, O. 202)

Unlike Western world rhetoric, which often emphasises spontaneity, the statement by the late Prime Minister of India reframed mass atrocity as equilibrium. "The big tree" metaphor becomes shorthand for justifiable revenge in communal discourse. It laid a political foundation for later populists to narrate violence not as an unlawful insolence but as an organic reaction.

This legacy matters because in today's digital world, the hate template carries more interpretive templates. When leaders or sectarian entrepreneurs frame minorities as threats, their audiences automatically draw on this memory: if they provoke us, we are justified in response. In this way, a single historic sentence becomes an endlessly reusable permission structure.

Monks and Mobilised Vigilantism

The most explicit modern example of populist hate documented in India concerns self-styled monks. These monks built their political and public brand. The core semantics of these self-styled monks' rhetoric mimics global populist grammar: minorities are framed out not as neighbours but as civilisational opponents. These speeches are not rumours; they exist on record, forming a base of many FIRs and attract judicial scrutiny. Yet unlike Orban or Bolsonaro, these monks' speeches travel almost immediately into news headlines, WhatsApp chats, YouTube shorts. Now a days police officials in numerous districts reported local mobs engaging in cow vigilantism. The mobs did not quote manifestos but speeches, "if they touch our cows, we will respond…… this is our country". The indirect support from local functionaries pushed the boundaries, transforming hate speech into de facto directive. PTI. (2022)

This is Indian innovation: speech escalates to policing substitution. Vigilante justice becomes governance.

Misappropriated Permission: Owaisi Brothers.

Akbaruddin Owaisi a populist leader occupies a paradoxical role within digital hate ecosystems. An almost 15-year-old short video, barely one minute, captures a speech in which he said, "We are 25 crores and you are 100 crores. Remove police for 15 minutes, Let's see who survives." . This clip is now circulated predominantly as a reel among young Muslim viewers; this fragment is consumed not as political rhetoric but as evidence of minoritarian aggression and resilience. The clip exemplifies, how digitally compressed speech generates affect, more efficiently than reasoned argument. The emotional high point, captured at the moment the audience erupts, becomes a memetic artefact that travels independently of intention or context (Udupa & Pohjonen, 2019. The paradox is accentuated by contrast. Asaduddin Owaisi, Member of parliament and his elder brother is known for his articulate, constitutionalist defence of minority dignity and rights, an oratory grounded in legal protections and democratic inclusion. Yet the viral 15-second Jibe overshadows such carefully reasoned speeches, demonstrating how digital infrastructures incentivise provocation and brevity over deliberative discourse. The leader no longer requires a structured argument: a moment of performative bravado substitutes for sustained ideological communication.

This mode of mobilisation is hybrid. Speech becomes both ammunition and content, a single phrase functions as a mnemonic device, a unit of political identity, and a cue for collective mobilisation. Even in periods when India's digital penetration was incomplete, such fragments operated as secondary vectors, enabling speech to circulate detached from its original narrative frame. Traditional legal categories are ill-equipped to address this phenomenon. India's criminal law targets intentional incitement, authorship, and explicit calls to violence, not remix, memeification, or adversarial clipping Platforms similarly struggle: moderation systems respond to virality, not context; acceleration triggers action, whereas interpretive harm remains invisible to automated systems The result is the emergence of digital shadow riots, wherein leaders speak in two languages simultaneously one literal, legally defensible, and one memetic, capable of producing hostility without overt instruction. In such regimes, accountability dissolves: the law sees a speech, while the digital public hears a battle cry.

Extremist Religious Entrepreneurs: The Dharam Sansad Model

The cleanest articulation of digital to physical escalation in India is observed in one of the religious conclaves known as Dharam Sansad. During 2020 to 2022, certain Hindu monks issued explicit genocidal calls against Muslims. The videos were not coded or whispered, but included direct appeals to procure weapons and prepare for ethnic cleansing. Participants recorded these speeches themselves and posted it on YouTube and telegram. These uploads circulated in WhatsApp group that organised self defence training sessions and roving vigilante patrols. They are not legislators nor executives, but they have digitally mediated moral authority. When they demand blood, it is interpreted as divine instruction because they are saints for their audience. This makes their rhetoric exceptionally dangerous. Their legitimacy bypasses democratic accountability entirely.

The Indian state's reluctance to intervene has been itself a permission signal. Arrests following Dharam Sansad's were delayed, cautious and often partial. Courts reprimanded organisers, but without consistent enforcement mechanism. Platform moderation removed some clips but many again resurfaced. The lesson transmitted to listeners was unmistakeable: the law does not punish your speech.

Caste Populism as Hate Infrastructure

Caste populism operates as a systematized hate infrastructure that leverages digital technologies to perpetuate discrimination and violence against marginalised communities (Teltumbde, 2018). This infrastructure transforms caste-based prejudice from isolated incidents into organised technologically enabled systems that normalize exclusion from society (Chakravarti, 2019). Digital platforms have altered the basis of how caste-based hate operates. Social media algorithms amplify inflammatory content, giving unprecedented reach to these narratives framed as defending this inhumane act (Raman, 2022).

Populist leaders exploits grievance politics, portraying affirmative action policies as attacks against merits, such rhetoric creates majoritarianism. Where democratic institutions are captured to enforce hierarchies rather than protection of rights (Appadurai, 2017).

Migrant Hate in North-East India

In North eastern India populist leader have targeted non tribals natives.

Their rhetoric emphasises on resource theft, cultural pollution and demographic erosion. Short clips of this rhetoric circulate in local digital networks often under the guise of community protection updates and broadcast. These types of groups compiles photographic dossiers and intimidate campaigns against migrant workers. These channels are invisible to regulators, as they are small but densely connected, forming retaliatory micro ecosystems, but this hate volcano erupts, it is synchronised, targeted and inorganic.

Regulation, Governance and Rights: India and European Comparative Framework

Law, Regulation and the Human Rights Dilemma

If digital hate were only a matter of speech legal systems could treat it under the familiar doctrines of expression, public order or defamation and incitement. The problem is that online hate is not confined to expression; but it is structurally entangled with patterns of criminal conduct- targeted assault, sexualised violence, force displacement, lynching, rioting. At this point question is not merely whether a sentence is "offensive"; it whether the digital infrastructure and populists' leaders are jointly producing conditions of foreseeable harm, which the state has an obligation to prevent.

International human rights law recognises that freedom of expression is not absolute. Under the international Covenant on civil and political rights (ICCPR), Article 19 protects expression, but article 20 explicitly requires states to prohibit "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (United Nations, 1966).

The difficulty is operational: when does rhetoric constitute "incitement" in an age where speech is fragmented, remixed and re-circulated beyond the author's direct control?

In the Indian context, domestic criminal law already contains provisions against this type of hate crimes, but enforcement has been selective, politically mediated and inconsistent. Populist leaders are rarely prosecuted, proceedings are slow and often inconclusive, whereas ordinary citizen expressing far less influential content sometimes faces swift and disproportionate sanctions. This asymmetry produces a legitimate crisis: the law seems to function as a instrument of power not neutral safeguard.

In India from a human rights perspective two obligations collide, firstly the duty to protect the freedom of speech and expression inc. political and religious speech, even when it is unsettling or disruptive. Secondly the duty to protect vulnerable group from targeted hostility, specially when it translates to violence. The first duty is often invoked rhetorically, while the second is neglected in practice. Populist narratives attempt to regulate hate speech as anti-majoritarian censorship, allowing perpetrators to recast themselves as victims of an authoritarian state, even where they benefit from state protection.

Digital platforms occupy middle ground and act as a catalyst; they control essential public infrastructure for deliberation. Where their algorithms demonstrably preferentially distribute content that encourages hate. Yet the existing regulatory frameworks have been slow to attach slow to attach concrete duties to the platforms. Another dimension is encryption and privacy. It protect citizens against state overreach and unlawful surveillance. From a rights-based perspective the question is not "should encryption be abolished"? but rather, how can states design context sensitive and rights respecting oversight mechanisms which allows early detection for systematic harm without establishing an infrastructure of total surveillance. It is normative challenge rather than a normative one.

Germany's NetzDG: A Controversial Pioneer

Germany's Network Enforcement Act (Netzwerkdurchsetzungsgesetz or NetzDG), enacted in 2017, represents one of the most significant attempts to regulate online hate speech through platform accountability (Library of Congress, 2021). The legislation requires social media platforms with more than two million users in Germany to remove the illegal content within 24 of receiving a complaint and other illegal content within seven days and failure to comply can result in fines up to €50 million (Ritzmann, 2019). The law references certain provisions of German criminal code including section 130 (Volksverhetzung—incitement to hatred), which prohibits advocacy that incited hatred against segment of population based on ethnic origin, religious, racial and nationality (PBS, 2022).

NetzDG has faced substantial criticism from multiple perspectives, human rights watch described the law as fundamentally flawed, as it creates vague, overbroad requirements that turn private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight (Human Rights Watch, 2018). The legislation has privacy concerns intensified when amendments required platforms to report removed illegal content including users IP addresses to federal police- a provision so controversial that president initially refused to sign it (Leerssen, 2020). Critics calls NetzDG a dangerous precedent for authoritarian regime throughout the world. Turkey, Russia, Malaysia have passed similar legislative explicitly referring the German model with far more repressive provisions (Forristal, 2017; Human Rights Watch, 2018). The Turkish version has been described as the worst version of Germany's NetzDG by the electronic frontier foundation implemented in a country with a second highest number of imprisoned journalists globally. (Forristal, 2017).

The EU Digital Services Act: Systemic Risk and Platform Duty

The European Union acknowledging the limitations of NetzDG, adopted the Digital Services act (DSA), 2022, which came into full effect in February 2024 (European Commission, 2025). This act represents a more comprehensive approach to the platform governance moving beyond the simple content removal mandate to address the systematic risks including algorithmic transparencies, and democratic oversights.

The DSA distinguishes between different categories of platforms. It is particularly stringent for very large online platforms with more than 45 million active users in European Union (Portaru, 2024). These "Very Large Online Platforms" (VLOPs) must conduct regular risk assessment, addressing negative effect on civic discourse, electoral processes, public security, gender-based violence, protection of public health and other serious negative consequences to persons, physical and mental health.

The DSA incorporated certain innovations, firstly the platforms must provide details report on content, moderation activities, terms of services, algorithmic functioning and advertisement practices. The amended framework includes appeals procedures for users whose content have been removed addressing the due process consensus raised by the NetzDG. Secondly users can flag content as illegal under the European union law or national law, with platforms requires to act expeditiously or face penalty up to the six percent of its global turn over.

However, DSA also faces its own critiques. The legislation relies on broad and contested terms like "disinformation", "misinformation" and "hate speech" that lack clear legal definition and are interpreted differently across the European member states (Portaru, 2024).

The framework "Brussels effect", whereby the European regulations shape global content moderation policies means that restrictive European standards may be applied worldwide and potential impacting speech in jurisdictions with different constitutional traditions (Bradford, 2020).

Towards a Normative Framework for Governing Digital Hate

The preceding analysis suggests that ordinary content moderation policies cannot address the depth of the problem. Removing individual posts or suspending accounts after major incidents treats hate as a series of isolated rule violations, instead of recognising it as a structural, intertwined with leadership, institutions and historical memories of impunities.

Leader Accountability: From Impunity to Responsibility

The most difficult part is to implement the leader's accountability. Populist leaders who repeatedly license hostility must be held to account. This does not mean that every controversial remark is criminalised. It means that where:

- Hate occurs in a context where groups are already under threat;
- Populist leaders use dehumanizing or demonising languages;
- Speech identifies specific targets;
- and is followed by pattern of violence, which clearly draw on rhetoric.

For the populist leaders, the issue is not of isolated outrages, but of systematic framework at lower inhibition against harm. Repeated rhetorical permission, under the condition of known risk should trigger heightened forms of scrutiny: Electoral sanctions, disqualification from office where appropriate and in grave cases, criminal liability in proportion to demonstrate impact.

Platform Duty of Care: From Neutral Carrier to Responsible Infrastructural entity.

The second pillar demand an explicit shift in how platform understand themselves. They cannot closely claim to be neutral conduct when:

- their algorithm amplifies emotional content;
- their features like (Reels, Status and Shorts) structurally favour fragmentary, decontextualized speech;
- their business models rely on maximising engagement regardless of model risk.

A duty of care framework would require platform to:

- audit their recommended system for systematic amplification of hate-laden content;
- develop real time disk monitoring for political and religious speech during the known flash point like communal anniversary, election controversial, controversial judicial decisions)
- introduce "friction" mechanisms in forwarding chains of potentially inflammatory content (e.g. limits on re-shares, context labels, click-through warnings);
- enable trusted flagger systems whereby credible civil society organisations and journalists can trigger accelerated moderation reviews for dangerous speech fragments from high-reach accounts.

Conclusion

The current debate globally situates human rights laws as an obstacle to security. encryption is blamed for concealing criminal activity. Freedom of speech is blamed for allowing extremists to thrive; free speech protections are blamed for allowing extremists to thrive; due process is blamed for "letting off" dangerous leaders. This paper advances the opposite view: a robust human rights framework is the only sustainable security strategy in the age of digital populism. If citizens believe that law is a tool to supress inconvenient speech rather than a shield against unjust harm, they will turn to extra-legal solutions. A human rights approach properly demand equality of enforcement, legality of proportionality, participatory regulation, international solidarity. The European experience demonstrates both the necessity and the difficulty of regulating digital platforms, whereas for India the path forward requires constitutional grounding, judicial oversight, sunset and review provisions.

The task is not to import European templates mechanically but to adapt governance principles of transparency, non-discrimination, accountability to realities. This requires acknowledging that each jurisdiction faces distinctive challenges. The sustainable solutions to these distinctive challenges require addressing the political economy of hate. They require rebuilding trust institutions that can mediate conflict fairly and requires collective commitment to the proposition that democracy is strengthened, not weakened, when it protects its most vulnerable members.

References

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209. https://doi.org/10.1207/S15327957PSPR0303_3

Benesch, S. (2014). *Dangerous speech: A proposal to prevent group violence.* World Policy Institute.

Berman, S. (2019). Populism is not fascism - but it could be a harbinger. *Foreign Affairs, 98*(6), 39–44.

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world.* Oxford University Press.

Chandrachud, A. (2017). *Republic of Rhetoric: Free Speech and the Constitution of India*. Penguin.

European Commission. (2025, January 20). *Codes of conduct under the Digital Services Act.* https://digital-strategy.ec.europa.eu/en/policies/dsa-codes-conduct

European Court of Human Rights. (2023). *Factsheet: Hate speech.* https://www.echr.coe.int/documents/d/echr/fs_hate_speech_eng

European Court of Human Rights. (n.d.). *Article 10: Hate speech—Key themes.* https://ks.echr.coe.int/documents/d/echr-ks/hate-speech

European External Action Service. (2025, June 16). *EU statement-High-level event: International Day for Countering Hate Speech.* https://www.eeas.europa.eu/delegations/un-new-york/eu-statement-high-level-event-international-day-countering-hate-speech_en

Forristal, C. (2017). German hate speech laws: Balancing freedom of expression and the fight against the incitement of hatred. *The Eagle Gazette, 8*(1). https://issuu.com/theeaglegazette/docs/eagle-vol8-issue1-final/s/13730278

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Human Rights Law Centre. (2018, April 6). *European Court of Human Rights finds hate speech not protected by freedom of expression.* https://www.hrlc.org.au/human-rights-case-summaries/2017/9/26/european-court-of-human-rights-finds-hate-speech-not-protected-by-freedom-of-expression

Human Rights Watch. (2018, February 14). *Germany: Flawed social media law.* https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Human Rights Watch. (2019). *Rainforest mafias: How violence and impunity fuel deforestation in Brazil's Amazon.* Human Rights Watch.

Leerssen, P. (2020). *A case study of Germany's NetzDG* [HAL Sciences Po Working Paper]. https://sciencespo.hal.science/hal-03586791/document

Leerssen, P. (2025, March 31). Strengthening the EU's digital landscape-Integration of the revised Code of Conduct on hate speech and the Code of Practice on Disinformation into the DSA. *MediaLaws.* https://www.medialaws.eu/strengthening-the-eus-digital-landscape-integration-of-the-revised-code-of-conduct-on-hate-speech-and-the-code-of-practice-on-disinformation-into-the-dsa/

Library of Congress. (2021, July 6). *Germany: Network Enforcement Act amended to better fight online hate speech.* https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/

Mudde, C. (2004). The populist zeitgeist. *Government and Opposition, 39*(4), 541–563. https://doi.org/10.1111/j.1477-7053.2004.00135.x

Pappas, T. S. (2019). *Populism and liberal democracy: A comparative and theoretical analysis.* Oxford University Press.

PBS. (2022, December 7). *Germany's laws on antisemitic hate speech and Holocaust denial.* https://www.pbs.org/wgbh/frontline/article/germanys-laws-antisemitic-hate-speech-nazi-propaganda-holocaust-denial/

Portaru, A. (2024). Is the EU's Digital Services Act compliant with the right to freedom of expression? *Oxford Human Rights Hub.* https://ohrh.law.ox.ac.uk/is-the-eus-digital-services-act-compliant-with-the-right-to-freedom-of-expression/

PTI. (2022, August 29). Haridwar Dharma Sansad case: Supreme Court directs hate speech accused to surrender by September 2. *The Hindu*. https://www.thehindu.com/news/national/other-states/haridwar-dharma-sansad-case-supreme-court-directs-hate-speech-accused-to-surrender-by-september-2/article65826343.ece

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr., W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW), 1–30. https://doi.org/10.1145/3392823

Ritzmann, A. (2019). *Fighting hate speech and terrorist propaganda on social media in Germany.* Program on Extremism, The George Washington University. https://extremism.gwu.edu/fighting-hate-speech-germany

Staff, O. (2020, November 2). Watch: Rajiv Gandhi's speech justifying 1984 anti-Sikh riots. *OpIndia*. https://www.opindia.com/2020/11/rajiv-gandhi-big-tree-falls-speech-indira-gandhi-assassination/

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media.* Princeton University Press.

Udupa, S., & Pohjonen, M. (2019). Extreme speech and digital cultures: An introduction. *International Journal of Communication, 13*, 3033–3043.

United Nations. (1966). *International Covenant on Civil and Political Rights.* United Nations Treaty Series, 999, 171.

Volokh, E. (2025, July 18). Hate speech and the European Court of Human Rights: Towards a principled approach. *Reason.* https://reason.com/volokh/2025/07/18/hate-speech-and-the-european-court-of-human-rights-towards-a-principled-approach/

Voorhoof, D. (2022). Conflicting conceptions of hate speech in the ECtHR's case law. *German Law Journal*, 23(9), 1193–1211. https://doi.org/10.1017/glj.2022.81

Positive obligations and online hate speech before the ECtHR: protection or chilling effects?

Francesca Cassano, University of Milan

1. Introduction

The rise of digital technologies and the global reach of online platforms have profoundly transformed the nature and impact of hate speech, creating an environment in which harmful expression can spread rapidly and often anonymously. Online spaces act as powerful amplifiers through mechanisms unique to virtual environments, including engagement-driven algorithms, the viral circulation of content, and the frequent absence of effective content moderation systems (Banks 2010; Ruotolo 2020). Although freedom of expression remains a cornerstone of democratic societies, there is also a pressing need to protect individuals and vulnerable groups from forms of expression that undermine their dignity, equality, and security. Nonetheless, the regulation of hate speech raises complex and contentious issues within the international human rights law framework.

At the international level, there is still no universally agreed definition; the concept remains shaped largely by soft-law instruments and interpretative practice. In its 2022 Recommendation on combating hate speech, the Council of Europe adopts a broad definition covering expressions that incite or justify violence, hatred or discrimination based on a wide range of personal characteristics (§2). The Recommendation also calls for hate speech to be categorised as requiring criminal law, civil or administrative sanctions, or alternative responses such as education and counter-speech (§3). Beyond the uncertainty surrounding its definition, there is also deep scholarly disagreement on how, and to what extent, hate speech should be regulated. Some scholars advocate a restrictive regulatory approach that emphasises the protection of human dignity and social cohesion (Matsuda 1993; Tsesis 2002; Waldron 2012), while others warn against over-regulation and its potential chilling effect on freedom of expression (Heinze 2017). These divergences are reflected in the varying thresholds and understandings adopted by international and regional human rights bodies. Consequently, the human rights law framework remains fragmented, caught between competing fundamental values and struggling to respond to the new challenges posed by this phenomenon in the online environment.

Against this critical backdrop, the European Court of Human Rights

("ECtHR" or the "Court") occupies a particularly prominent and, at times, controversial position in the regulation of hate speech. The European Convention on Human Rights ("ECHR" or the "Convention") contains no explicit provision on hate speech, yet it has progressively developed the most comprehensive bodies of jurisprudence on the matter worldwide. Notably, a recent development in the Court's case law revealed its propensity to address and recognise States' positive obligations regarding hate speech by applying Article 8 alone or in conjunction with Article 14 of the ECHR. This remarks a conceptual turning point, as hate speech is no longer framed solely in terms of permissible limitations on freedom of expression under Article 10 paragraph 2 of the ECHR, but also as a matter of State duty to protect individuals and groups from its negative effects. However, this development has also raised important concerns about the exact nature, extent, and boundaries of such positive obligations.

To this end, the present study will first examine the ECHR's legal framework and the evolution of the Court's jurisprudence in the absence of an explicit hate speech provision. It will then analyse the content and scope of States' positive obligations to protect individuals and groups from online hate speech, as developed in the Court's recent case law under Articles 8 and 14 of the Convention, with particular attention to the 2025 judgments Minasyan and Others v. Armenia and Ilareva and Others v. Bulgaria. A comparative section will subsequently evaluate the ECtHR's approach alongside that of relevant monitoring bodies within the United Nations ("UN") framework, in order to determine whether the Strasbourg system mirrors, diverges from, or advances prevailing international practice on hate speech regulation. This research's overarching aim is to critically evaluate the ECtHR's contribution to the international regulation of hate speech as the most widespread judicial framework in this field, and the assessment of whether the broad recognition of positive obligations may generate uncertainty and undue restrictions on freedom of expression.

2. Bridging a normative gap: The ECtHR's jurisprudential development on hate speech

As previously mentioned, unlike other international and regional human rights instruments, the ECHR lacks a specific hate speech provision. Within the UN Treaty Body system, two pivotal provisions are generally regarded as the main international norms addressing hate speech. These are Article 20 of the International Covenant on Civil and Political Rights ("ICCPR") and Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination ("ICERD"). Although neither of these articles explicitly refers to the term "hate speech", they have traditionally been interpreted as forming the normative basis of the contemporary regulation of this phenomenon within the framework of human rights law (Farrior 1996). At a regional level, the ECHR also differs from the American Convention on Human Rights ("ACHR"), which includes a specific provision prohibiting certain forms of expression in Article 13, paragraph 5.

This absence can also be understood in light of the historical context in which the Convention was drafted. Adopted in 1950, just two years after the Universal Declaration of Human Rights (1948), which also contains no explicit reference to hate speech or specific limitation of the right to freedom of expression, the ECHR reflects an earlier stage in the international regulation of expression. Despite presenting this normative gap, however, the ECtHR has progressively become a prominent forum for adjudicating and developing hate speech jurisprudence, demonstrating an increasing commitment to addressing this challenge. Furthermore, its sustained attention to hate speech reflects the evolving interpretative practice of the Court to preventing the re-emergence of totalitarian ideologies rooted in intolerance, exclusion, and incitement to hatred. Conversely, the ECtHR has not adopted an exhaustive definition of hate speech in its case law, instead approaching the concept on a case-by-case basis. While this flexible approach enables the Court to consider contextual and societal specifics, it has also been criticised for lacking clarity and predictability especially when legal responses are required (McGonagle 2013, 11).

Traditionally, hate speech cases have been addressed by the ECtHR through the alternative application of Articles 10 paragraph 2 and 17 of the Convention. The former guarantees freedom of expression and sets out the conditions under which it may be lawfully restricted by the States, while the latter functions as the Convention's anti-abuse clause. These provisions have primarily been invoked in cases where applicants have alleged a violation of their right to freedom of expression. In other words, when applying these provisions, the ECtHR is usually required to verify whether a State has breached its negative obligation not to interfere with the right to freedom of expression by illegitimately invoking hate speech restrictions to protect the rights of others. While this study focuses on cases where applicants were victims of hate speech and the subsequent State failure to protect them, an examination of how the Court interprets and applies these articles is essential to gather a clearer understanding of its overall approach to hate speech.

From the Court's case law, it is evident that a hierarchical approach is employed when applying Articles 17 and 10(2) to hate speech cases. In other words, the ECtHR distinguishes between the most severe and less severe forms of hate speech. When invoking Article 17, the Court effectively excludes the alleged expression from the protection of Article 10 altogether, on the basis that the Convention cannot justify actions that destroy its underlying values. In such circumstances, the application is declared inadmissible ratione materiae, without any substantive examination as to whether the interference with freedom of expression was lawful or necessary (Castellaneta 2017). Conversely, when relying on Article 10(2), the Court acknowledges that a given expression is protected by Article 10, yet it may still be subject to the legitimate restrictions provided in the second paragraph. Thus, in these cases, the Court examines whether the interference was prescribed by law, pursued a legitimate aim and was necessary in a democratic society, as well as whether it was proportionate to the aim in question. Through this provision, the ECtHR has generally dealt with expressions that, while offensive, provocative or discriminatory, do not reach the level of severity required for exclusion under Article 17. This hierarchical distinction is clearly reflected in the Court's judgements.

Accordingly, Article 17 has been invoked in relation to several instances of hate speech, including Holocaust denial, as well as cases of incitement to hatred on the basis of ethnicity, religion, or sexual orientation. In Norwood v. the United Kingdom (2004), displaying a poster depicting the Twin Towers in flames alongside a statement calling for the expulsion of Muslims from Britain was deemed an attack on a religious community, equating it entirely with terrorism and thus contravening the fundamental principles of tolerance, social peace and non-discrimination enshrined in the Convention. More recently, in the 2023 Lenis v. Greece case, the Court found that online statements by a senior Orthodox Church official denying the humanity of LGBTIQ+ people and inciting violence against them constituted the most serious form of hate speech (Johnson 2024). This was particularly significant given the applicant's position of influence and the substantial harm that could be caused, thus establishing the applicability of Article 17 (Lenis v. Greece, §49, 50, 51). Conversely, it has also been argued that excluding certain forms of expression from the protection of Article 10 altogether is unnecessary for safeguarding democratic principles, and that it prevents a proper proportionality assessment and limits the development of consistent, transparent standards for evaluating hate speech (Cannie and Voorhoof 2011).

Whereas, when determining whether an interference with freedom of expression meets the conditions set out in Article 10(2), the ECtHR generally considers a range of factors such as the content and context of the speech, the intent of the speaker, and the likelihood of harm, thereby adopting a context-sensitive approach. Notably, the Court was the first adjudicatory body to extend these principles to the online environment, recognising the particular risks posed by the digital dissemination of hate speech and the responsibilities of online platforms (see, for example, Delfi AS v. Estonia, 2015, Sanchez v. France, 2023 and Google LLC and Others v. Russia, 2025). For instance, in the landmark Sanchez v. France case, the Court ruled that the criminal fine issued to the applicant, a politician, did not contravene Article 10, as he had neglected to swiftly remove third-party comments containing hate speech from his Facebook "wall" (§189).

On the other hand, the ECtHR's application of Article 10(2) has raised concerns regarding the breadth of discretion afforded to States under the margin of appreciation doctrine, as States are frequently granted substantial deference when defining the scope of expression and assessing the necessity of particular restrictions (Mchangama and Alkiviadou 2021; Sottiaux 2022). The Court has indeed adopted a relatively broad approach to upholding criminal sanctions for hate speech, which is more permissive than the standards applied by the UN

Human Rights Committee or other regional bodies (Clooney and Neuberger 2024, 63). For instance, the Court has consistently affirmed the compatibility of criminal penalties for hate speech with the right to freedom of expression in cases involving Holocaust denial, an approach that contrasts with the Human Rights Committee's General Comment No. 34, which states that laws penalising opinions on historical facts are incompatible with the Covenant's protections (§49). This distinctive position reflects the ECtHR's view that hate speech threatens not only the individuals or groups it targets, but also the fundamental principles on which society is based, such as democracy and social inclusiveness.

More recently, as anticipated, the Court's jurisprudence has also begun to address cases brought by alleged victims of hate speech who claim that States had failed to take adequate protective or preventive measures. In particular, the ECtHR's progressively evolved to extend protection to victims of hate speech under Article 8 of the Convention (right to respect for private life), both alone and in conjunction with Article 14 (prohibition of discrimination), and, in exceptional circumstances, under Article 13 (right to an effective remedy). The Court first engaged with this line of reasoning in the 2012 case of Aksu v. Turkey, in which the applicant argued that certain publications funded by the State contained anti-Roma expressions, amounting to a violation of his rights under Articles 8 and 14. Although the Court ultimately found no violations, this case marked a pivotal moment in recognising that hate speech could infringe the right to private life and that States have a positive obligation to prevent and respond to such harm. The Aksu judgment thus laid the groundwork for a gradual clarification and expansion of States' duties to protect individuals and groups from hate speech, particularly where it intersects with systemic discrimination or prejudice.

However, in the absence of a specific provision in the ECHR that explicitly addresses States' positive duties with regard to hate speech, it remains challenging to identify the precise content of these obligations ex ante. Some scholars have questioned the ECtHR's approach, suggesting that it may place excessive limitations on the right to freedom of expression due to the uncertainty surrounding the exact scope of these obligations (Alkiviadou 2025, 178). Nevertheless, the Court's case-by-case assessments of the adequacy and proportionality of domestic responses to hate speech incidents have progressively delineated the parameters of States' positive duties. The following section examines this jurisprudential approach in detail, exploring how the ECtHR has defined States' positive obligations under articles 8 and 14, and considering whether this expanding protective framework enhances victims' rights or risks unduly restricting freedom of expression.

3. Positive obligations under Articles 8 and 14 of the ECHR: protection or restriction?

Firstly, regarding the concept of positive obligations: traditionally, international human rights law distinguishes between negative obligations, which require States to refrain from interfering with the enjoyment of fundamental rights, and positive obligations, which compel them to take active measures to ensure the effective protection of those rights (Shue 1997; Mazzeschi 2008; Shelton and Gould 2013). This distinction is employed when determining State responsibility, as it helps to identify the type of State conduct, thus whether a breach results from an act or from an omission. As has been contended, omissions are more than mere "inaction", acquiring legal relevance only when a duty to act exists and remains unfulfilled, and their significance can only be assessed in light of the content and scope of that duty (Crawford 2013, 218).

Within the framework of the ECHR, this distinction is of particular importance, presenting complex interpretative challenges (Lavrysen 2016; Stoyanova 2023). Article 1 of the ECHR requires States not only to respect, but also to secure the rights enshrined in the Convention. These rights are formulated in broad and general terms, leaving significant room for judicial interpretation. In line with its well-established doctrine that the Convention is a "living instrument", the ECtHR has interpreted these provisions evolutively, adapting their meaning to contemporary social realities and emerging threats to human rights (Letsas 2012). Accordingly, the Court has progressively recognised that the protection afforded by the ECHR entails not only duties of abstention, but also positive duties requiring States to adopt different kinds of measures to safeguard individuals from interferences by both public authorities and private actors (see, for example, these early judgements: Belgian Linguistic Case, 1968 and Marckx v. Belgium, 1979). This dynamic interpretative approach has enabled the Court to extend the scope of the Convention's rights to new contexts, including the growing phenomenon of online hate speech.

In light of this interpretative evolution and to clarify the content and scope of States' positive obligations, this analysis will begin with Article 8 of the Convention and the obligations arising therefrom with regard to the States' duty to protect vulnerable groups and individuals from hate speech. Firstly, regarding the applicability of the right to respect for private and family life to hate speech cases, the Court has clarified that, even if the unlawful treatment in question does not reach the level of severity required to fall within the scope of Article 3 (prohibition of torture and inhuman or degrading treatment), it may nonetheless engage Article 8 (R.B. v. Hungary § 79, 2016; Király and Dömötör v. Hungary, 2017, §42; Association ACCEPT and Others v. Romania §66, 2021). This is because any stereotyping or denigration of a group, when reaching a certain level of intensity, is capable of affecting the group's sense of identity, as well as the self-worth and self-confidence of its members, thereby interfering with their private life within the meaning of Article 8 (Aksu v. Turkey, 2012, § 58; Király and Dömötör v. Hungary, 2017, § 41). When considering whether this level of severity has been reached, the Court takes into account factors such as the group's historical vulnerability, the content of the negative stereotype in question, and the context and form in which it has been conveyed.

Recent case law has further clarified the notion of victim status in this context.

Notably, in the aforementioned cases of Minasyan and Others v. Armenia and Ilareva and Others v. Bulgaria (2025), the Court found a violation of Article 8 even though the applicants were not directly part of historically marginalised groups. Instead, they were targeted because of their association with such groups, in their capacity as activists and human rights defenders. In Minasyan, the applicants were attacked for both their activism and their perceived sexual orientation, as well as their association with the LGBTIQ+ community (§54). Similarly, in Ilareva, the Court found that the threatening and denigrating statements directed at the applicants on public Facebook pages, motivated by their activism in defence of refugees, constituted an affront to their psychological integrity and dignity, thus falling within the scope of Article 8 (§105, 116).

Once the applicability of Article 8 has been established, the Court will typically verify the existence of an effective legal framework that enables vulnerable groups and individuals to assert the rights protected under Article 8. Although States have discretion over the specific measures to adopt, the Court examines whether these measures are reasonable, effective and balanced with the public interest of safeguarding freedom of expression. Regarding the latter, the ECtHR has also clarified that the imposition of criminal sanctions for hate speech should be regarded as an extrema ratio measure, thus to be applied only when strictly necessary, and implemented in a manner consistent with the State's broader obligation to ensure effective protection of the rights encompassed under Article 8. Therefore, for ensuring an adequate legal framework, States may instead fulfil their obligations through alternative mechanisms, such as administrative or civil remedies, provided that they offer effective and sufficient protection. For instance, in Minasyan and Others v. Armenia, the Court ruled that, while the domestic civil law framework did not specifically address instances of discrimination on the basis of sexual orientation, it was, at least in theory, capable of providing protection to the applicants from encroachment on various aspects of their private life within the meaning of Article 8, including from homophobic hate speech (§64).

Furthermore, as previously mentioned, Article 8 can be invoked either alone or in conjunction with Article 14, which prohibits discrimination. With regard to the latter article, States' positive obligations primarily relate to the duty to conduct effective investigations. More broadly, the obligation to investigate constitutes a procedural positive obligation, and it has been identified, also by the Court, as an obligation of means rather than of result (Ilareva and Others v. Bulgaria, § 135). The Court's approach to this issue largely mirrors its reasoning in hate crime cases, with the distinction that in such cases Article 14 is typically applied in conjunction with Article 3, rather than Article 8. As in hate crime jurisprudence, the principle of non-discrimination acts as a lens through which the ECtHR underscores the State's additional duty to take all reasonable steps to uncover any bias motive and to determine whether prejudice or hatred played a role in the events (Pressacco 2024). In other words, where there is evidence that the impugned statements were motivated by, or had the effect of, discrimination, the assessment of an investigation's effectiveness goes beyond the traditional requirements of promptness, diligence, and impartiality, demanding also that States meet a higher standard in responding to alleged bias-motivated incidents.

In the case of Ilareva and Others v. Bulgaria, concerning a series of Facebook posts containing death threats and racially motivated insults against refugee rights activists, the Court provided further on the State's duty to conduct effective investigations in cases of online hate speech. The ECtHR criticised the domestic authorities for failing to pursue available investigative avenues that could have identified the perpetrators, such as requesting the collection of digital traffic data linked to the IP addresses of those who had posted the threats (§119, 120). Out of the eleven individuals involved, only two were identified and just one was questioned (§121). The prosecuting authorities had also unjustifiably downplayed the seriousness of the threats, dismissing them as expressions of personal opinion rather than discriminatory crimes. The Court held that these omissions resulted in impunity for the perpetrators and left the applicants unprotected, thereby encouraging similar conduct to continue and escalate (§122).

In Beizaras and Levickas v. Lithuania (2020), the Court considered a similar instance of State protection failing to address online hate speech, this time in the form of homophobic comments posted under a photograph of two men kissing. This judgment is significant because in addition to Articles 14 and 8, it is the only instance in hate speech jurisprudence where the Court also found a violation of Article 13. While Article 13 is not typically examined separately when a violation of Article 14 has been established, the Court ruled that the discriminatory attitudes of the domestic authorities in this case had rendered existing remedies ineffective and warranted independent scrutiny. The Court also found that the authorities had consistently failed to investigate hate speech cases involving sexual minorities (§155). The Court therefore noted a pattern of institutional prejudice, including instances where domestic courts explicitly described hate speech investigations as a "waste of time and resources" (§23). Consequently, in light of the risk of rendering national hate speech provisions ineffective and of demonstrating a failure by law enforcement to take bias-motivated offences seriously, the Court concluded that there had been a violation of both the applicants' rights to an effective remedy and their right to non-discrimination in conjunction with their right to private life.

In conclusion, this jurisprudence reflects the ECtHR's increasing commitment to ensuring that States fulfil their positive obligations to protect individuals and groups from hate speech, particularly in light of systemic failures or insufficient responses by domestic authorities. The examined cases reveal a clear effort by the Court to address structural deficiencies that perpetuate impunity for bias-motivated abuses, and to affirm the need to effectively protect victims online. However, the ECtHR's expanding interpretation of positive obligations has raised legitimate concerns about its potential impact on freedom of expression and the risk of excessive State interference in the public sphere. To mitigate these risks and promote greater legal certainty, the Court could further clarify the parameters governing the seriousness and intensity required for hate speech to engage State responsibility under Articles 8 and 14. More precise and consistent criteria, particularly with regard to the exceptional circumstances warranting criminal sanctions as opposed to civil or administrative measures, would enhance both foreseeability and compliance. Therefore, while a case-by-case assessment is essential due to the contextual nature of online hate speech, clearer guidance from the Court could help to ensure a balanced approach, strengthening protection against hate speech while avoiding over-regulation and undue restrictions on freedom of expression.

4. Converging or diverging paths? The ECtHR and UN Bodies on States' duty to protect from hate speech

In considering the UN treaty bodies system, the analysis will focus on the aforementioned provisions contained in Articles 20 of the ICCPR and 4 of the ICERD. Article 20 of the ICCPR requires States Parties to prohibit by law any propaganda for war as well as any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. Article 4 of the ICERD goes further by obliging States to criminalise certain forms of expression and conduct that promote racial hatred and discrimination. It requires States to declare the dissemination of ideas based on racial superiority or hatred, as well as any incitement to racial discrimination, as offences punishable by law, and to criminalise organisations and activities that promote or incite such discrimination. Taken together, these provisions represent the only instances within the UN treaty framework in which States are explicitly required to prohibit or criminalise hate speech. The concluding part of this study will therefore examine these two provisions, paying particular attention to how their respective monitoring bodies have interpreted and clarified the scope and content of the positive obligations they entail.

Starting with the interpretation and application of Article 20 by the Human Rights Committee ("HRC"), i.e. the monitoring body of the ICCPR, it should be noted that implementing this provision has proven to be particularly challenging for States, primarily due to definitional ambiguities and interpretative uncertainty surrounding its key terms. Concepts such as "advocacy", "hatred", and "hostility" lack precise legal definitions, and the element of incitement has generated substantial debate regarding the threshold of intent and causation required to trigger State responsibility (Elbahtimy 2021, 31). These difficulties mirror the wider challenge of defining hate speech under international law. Conversely, the HRC has not yet adequately explained the extent of the positive obligations incumbent upon States with regard to the norm in question. For instance, no general comment has been adopted regarding the interpretation of Article 20. In fact, General Comment No. 11 (1983) does little to clarify the content and scope of this provision; it merely offers a brief explanation of its compatibility with the right to freedom of expression enshrined in Article 19 of the ICCPR, essentially encouraging States to comply. The HRC's limited case law further reflects this cautious approach. To date, the HRC has never found a violation of Article 20, and only two individual communications have been submitted by victims of hate speech (Maria Vassilari et al. v. Greece, 2009 and Rabbae v. the Netherlands, 2016). In Maria Vassilari, the part of the communication concerning the alleged violation of Article 20(2) was declared inadmissible for lack of substantiation (§6.5), while the only case examined on the merits was Rabbae v. the Netherlands, where the applicants claimed that the State had failed to protect them from anti-Muslim statements made by a politician. On that occasion, while recognising the right of individuals and minority groups to be protected from incitement, the only consideration made by the HRC regarding the positive obligations under Article 20(2) was that States are not required to secure the conviction of individuals charged with such offences, but rather to ensure the existence of an adequate legislative framework (§10.7).

In contrast to the HRC's cautious approach under Article 20 of the ICCPR, the Committee on the Elimination of Racial Discrimination ("CERD") has adopted a far more proactive and detailed stance in clarifying the scope and nature of States' obligations under Article 4 of the ICERD. This interpretative and monitoring activity has been pivotal in turning the provision's broad language into a more concrete legal framework for combating racist hate speech. In this context, the Committee's General Recommendation No. 35 on Combating Racist Hate Speech (2013) is a crucial interpretative instrument. In this document, the CERD explicitly acknowledges that Article 4 entails obligations to prevent racist hate speech by introducing specific criminal offences, as well as ensuring their effective implementation. The Recommendation emphasises that criminalisation should be reserved for the most serious instances of hate speech, reaffirming the extrema ratio nature of criminal law measures. However, it also emphasises that States must adopt preventive, educational and policy measures to address the root causes of prejudice and intolerance, in addition to enacting legislation. The Committee's interpretative efforts have been further consolidated through its case law. For instance, in the most recent individual communication, the applicant, a human rights defender of African descent, claimed that publishing discriminatory images of him and his ethnic group during a public exhibition violated Articles 4 of the ICERD (Momodou Jallow v. Denmark, 2023). Although an investigation had been formally conducted and considered effective, the CERD found that the decision of the domestic authorities to dismiss the criminal proceeding was not an adequate or proportionate response, given that the images in question promoted ideas of racial superiority and thus fell within the scope of Article 4 (§7.13). The CERD held therefore that Denmark had failed to take the necessary measures to ensure the applicant's effective protection, thereby breaching its positive obligation to effectively implement measures against racist hate speech.

Finally, with regard to the legal nature of these provisions, it has been argued that they embody not only the aforementioned positive obligations, but also a correlative right to be effectively protected from incitement to hatred (Temperman 2019; Elbahtimy 2021). In this sense, the provisions operate on two complementary levels: they impose enforceable duties on States and simultaneously recognise a substantive right of vulnerable individuals and groups to live free from incitement to hatred. Therefore, both the ECtHR and the UN treaty bodies share the common objective of ensuring effective protection against hate speech and of framing States' duties in positive terms. However, their approaches diverge significantly. While the UN system offers a normative basis through Articles 20 of the ICCPR and 4 of the ICERD, the ECtHR has developed its standards through case law without an explicit textual basis.

Nevertheless, the ECtHR's "living instrument" approach has enabled it to extend protection online and to other vulnerable groups, including those targeted for their sexual orientation or gender identity, beyond traditional grounds such as race, nationality or religion. This evolution highlights the Court's dynamic role, but it also reveals a structural gap: the lack of clear definitions, scopes and thresholds for hate speech continues to create uncertainty and ambiguity within the ECHR framework.

5. Concluding remarks

In conclusion, the ECtHR has played a pivotal role in shaping the modern understanding of hate speech as well as the States' positive obligations to protect individuals and groups from its negative consequences. Despite there being no explicit provision for this in the Convention, the Court has progressively developed a coherent body of jurisprudence. In particular, under Articles 8 and 14 the Court explicitly recognises the necessity of State action to prevent and respond to hate-motivated expression. This development is especially significant in the context of online communication, where the Court has recognised the increased harm and rapid spread of digital hate speech, and the potential liability of online platforms in addressing it. Notably, the ECtHR is the only international judicial body that has directly adjudicated cases concerning hate speech disseminated through social media comments and has assessed the potential liability of internet intermediaries for failing to remove or properly moderate such content in this context.

At the same time, the Court's increasing focus on positive obligations mirrors a wider change in the international discourse: in an era where the rapid spread and reach of hate speech can endanger democratic values and human dignity, a purely liberal approach is no longer feasible. This evolution also illustrates the Court's gradual move from a predominantly negative conception of State duties towards a more substantive understanding of human rights protection, one that demands proactive engagement by States in countering hate speech. However, the necessity of limiting the effects of hate speech must always be balanced with the fundamental right to freedom of expression. In this regard, the ECtHR could strengthen its contribution by not only encouraging legislative and judicial domestic responses, but also educational and preventive measures, in line with the CERD Committee's interpretative guidance under Article 4 of the ICERD.

Overall, when considering States' positive obligations within the broader human rights framework, certain convergences emerge. Both universal and regional systems now recognise the duty to establish effective legal and institutional frameworks to prevent, investigate and proportionally sanction hate speech, thereby ensuring the protection of dignity and equality. However, differences persist: the UN framework provides explicit legal obligations, albeit with limited scope, whereas the ECtHR offers broader, evolving protection through judicial interpretation, albeit at the expense of legal clarity. This creates a landscape of both complementarity and fragmentation, which highlights the need for greater coherence and shared standards in defining States' duties to protect against hate speech in the digital age. Ensuring such coherence will ultimately be essential to transforming the protection against online hate speech from a fragmented legal aspiration into a concrete, enforceable human right.

References

Alkiviadou, N. (2025). *Hate speech and the European Court of human rights*. Routledge.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3), 233–239. doi.org/10.1080/13600869.2010.522323.

Cannie, H., & Voorhoof, D. (2011). The Abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection? *Netherlands Quarterly of Human Rights*, 29(1), 54–83. doi.org/10.1177/016934411102900105.

Castellaneta. M. (2017). La Corte europea dei diritti umani e l'applicazione del principio dell'abuso del diritto nei casi di 'hate speech'. *Diritti umani e diritto internazionale*, 3, 745–751. doi.org/10.12829/88608.

Clooney, A., & Neuberger, D. E. (2024). *Freedom of speech in international law.* Oxford university press.

Crawford, J. (2013). *State Responsibility: The General Part*. Cambridge: Cambridge University Press.

Elbahtimy, M. (2021). *The Right to Protection from Incitement to Hatred: An Unsettled Right*. Cambridge: Cambridge University Press.

Farrior, S. (1996). Molding the Matrix: The Historical and Theoretical

Foundations of International Law concerning Hate Speech, *Berkeley Journal of International Law*, 14(1) doi.org/10.15779/Z38J34B.

Heinze, E. (2017). *Hate speech and democratic citizenship* (First published in paperback). Oxford University Press.

Johnson, P. (2024). Homophobic Hate Speech and Article 17 ECHR: The Evolving Approach of the European Court of Human Rights. *SSRN* Electronic Journal. doi.org/10.2139/ssrn.4801753.

Lavrysen, L. (2016). *Human Rights in a Positive State: Rethinking the Relationship between Positive and Negative Obligations under the European Convention on Human Rights* (1st edn). Intersentia. doi.org/10.1017/9781780685311.

Letsas, G. (2012). The ECHR as a Living Instrument: Its Meaning and its Legitimacy. *SSRN* Electronic Journal. doi.org/10.2139/ssrn.2021836.

Matsuda, M. J. (1993). *Words that wound: Critical race theory, assaultive speech, and the First Amendment*. Westview Press.

Mazzeschi, R. P. (2008). Responsabilité de l'État pour violation des obligations positives relatives aux droits de l'homme (Volume 333). In *The Hague Academy Collected Courses Online / Recueil des cours de l'Académie de La Haye en ligne*. Brill | Nijhoff. doi.org/10.1163/1875-8096_pplrdc_A9789004172845_02.

McGonagle, T., (2013). *The Council of Europe against Online Hate Speech: Conundrums and Challenges.* Expert Paper. Belgrade: Ministry of Culture and Information, https://rm.coe.int/168059bfce.

Mchangama, J., & Alkiviadou, N. (2021). Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb? *Human Rights Law Review*, 21(4), 1008–1042. doi.org/10.1093/hrlr/ngab015.

Pressacco, L. (2024). Azione penale e tutela dei diritti fondamentali. L'obbligo di svolgere investigazioni effettive e la persecuzione degli "hate crimes". *Diritto e Società Plurale: Questioni Aperte*, 155–170.

Ruotolo G. M. (2020). A Little Hate, Worldwide! Di libertà d'opinione e discorsi politici d'odio on-line nel diritto internazionale ed europeo. *Diritti umani e diritto internazionale*, 2, 549–582. doi.org/10.12829/97968.

Shelton, D. (Ed.). (2013). *The Oxford Handbook of International Human Rights Law* (1st edn). Oxford University Press. doi.org/10.1093/law/9780199640133.001.0001.

Shue, H. (1997). *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy: 40th Anniversary Edition*. Princeton University Press. doi.org/10.1515/9780691200835.

Sottiaux, S. (2022). Conflicting Conceptions of Hate Speech in the ECtHR's Case Law. *German Law Journal*, 23(9), 1193–1211. doi.org/10.1017/glj.2022.81.

Stoyanova, V. (2023). *Positive Obligations under the European Convention on Human Rights: Within and Beyond Boundaries*. Oxford University Press.

Temperman, J. (2019). The International Covenant on Civil and Political Rights and the "Right to be Protected against Incitement". *Journal of Law, Religion and State*, 7(1), 89–103. doi.org/10.1163/22124810-00701005.

Tsesis, A. (2002). *Destructive Messages: How Hate Speech Paves the Way For Harmful Social Movements*. New York University Press. doi.org/10.18574/nyu/9780814784297.001.0001.

Waldron, J. (2012). *The harm in hate speech*. Harvard university press.

<u>*Case law*</u>

CERD, communication No. 62/2018, *Momodou Jallow v. Denmark*, UN Doc. CERD/C/108/D/62/2018 (12 January 2023).

ECtHR, *Case 'Relating to Certain Aspects of the Laws on the Use of the Languages in Education in Belgium' v. Belgium*, Application no. 1474/62 (23 July 1968).

ECtHR, *Marckx v. Belgium*, Application no. 6833/74 (13 June 1979).

ECtHR, *Norwood v. the United Kingdom*, Application no. 23131/03 (16 November 2004).

ECtHR, *Aksu v. Turkey*, Applications nos. 4149/04 and 41029/04 (15 March 2012).

ECtHR, *Delfi AS v. Estonia*, 2015, Application no. 64569/09 (16 June 2015).

ECtHR, *R.B. v. Hungary*, Application no. 64602/12 (12 April 2016).

ECtHR, *Király and Dömötör v. Hungary,* Application no. 10851/13 (17 January 2017).

ECtHR, *Beizaras and Levickas v. Lithuania*, Application no. 41288/15 (14 January 2020).

ECtHR, *Association ACCEPT and Others v. Romania*, Application no. 19237/16 (1 June 2021)

ECtHR, *Sanchez v. France*, Application no. 45581/15 (15 May 2023).

ECtHR, *Lenis v. Greece,* Application no. 47833/20 (31 August 2023).

ECtHR, *Minasyan and Others v. Armenia,* Application no. 59180/15 (7 January 2025).

ECtHR, *Google LLC and Others v. Russia*, Application no. 37027/22 (8 July 2025).

ECtHR, *Ilareva and Others v. Bulgaria,* Application no. 24729/17 (9 September

2025).

Human Rights Committee, Communication No. 1570/2007, *Maria Vassilari et al. v. Greec*e, UN Doc. CCPR/C/95/D/1570/2007 (29 April 2009)

Human Rights Committee, Communication no. 2124/2011, *Rabbae, A.B.S and N.A v. The Netherlands*, UN Doc. CCPR/C/117/D/2124/2011 (18 November 2016).

International documents

Council of Europe, European Convention on Human Rights, as amended by Protocols Nos. 11, 14 and 15, ETS No. 005, 4 November 1950, https://www.refworld.org/legal/agreements/coe/1950/en/18688.

Council of Europe, Recommendation of the Committee of Ministers to member States on 'combating hate speech' (Adopted by the Committee of Ministers on 20 May 2022 at the 132nd Session of the Committee of Ministers), CM/Rec(2022)16, https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680a67955%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}.

UN Committee on the Elimination of Racial Discrimination (CERD), *General recommendation* No. 35: Combating racist hate speech, CERD/C/GC/35, 26 September 2013, https://www.refworld.org/legal/general/cerd/2013/en/101142.

UN General Assembly, Universal Declaration of Human Rights, 217 A (III), 10 December 1948, https://www.refworld.org/legal/resolution/unga/1948/en/11563.

UN General Assembly, *International Covenant on Civil and Political Right*s, United Nations, Treaty Series, vol. 999, p. 171, 16 December 1966, https://www.refworld.org/legal/agreements/unga/1966/en/17703.

UN General Assembly, International Convention on the Elimination of All Forms of Racial Discrimination, United Nations, Treaty Series, vol. 660, p. 195, 21 December 1965, https://www.refworld.org/legal/agreements/unga/1965/en/13974.

UN Human Rights Committee (HRC), CCPR General Comment No. 11: Article 20 Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred, 29 July 1983, https://www.refworld.org/legal/general/hrc/1983/en/33808.

UN Human Rights Committee (HRC), General comment no. 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, https://www.refworld.org/legal/general/hrc/2011/en/83764.

Part III

Gendered Hate and Online Misogyny: Narratives, Movements, and Countervoices

Emergency Meeting 63: Extreme Misogyny, Andrew Tate, and the Stoking of Collapse

Elizabeth Pearson, Royal Holloway, University of London

Introduction

On 29 July 2024 17-year-old Axel Rudakubana arrived at the Hart Space Studio in Southport England, where a Taylor Swift-themed dance class for young children was taking place. Rudakubana, armed with a knife, proceeded to carry out an attack that would leave three children dead: seven-year-old Elsie Dot Stancombe, six-year-old Bebe King, and nine-year old Alice Da Silva Aguiar. Eight other children were injured, along with two adults (The Crown Prosecution Service 2025). The attack provoked an outpouring of national sadness and outrage; it also gained widespread traction across social media. Many individuals speculated on the identity of the attacker, who was not initially named (Watling 2024). Within hours of the incident, misinformation spread online suggesting the attacker was an illegal immigrant who had arrived by small boat, and whose name was 'Ali Al Shakti' (ISD 2024). Within a few days of the attack, 'far right' race riots had broken out in more than 20 towns and cities across England, rioters looting shops, and attacking mosques and asylum hostels.

Among those actively posting online to inflame the situation were anti-immigration Reform politician Nigel Farage, who questioned whether 'the truth was being withheld' (Dodd et al. 2024); anti-Islam Street Movement 'English Defence League' founder turned self-described 'journalist' Stephen Yaxley-Lennon aka Tommy Robinson, who posted an online video on the topic of 'resistance'; and online manfluencers Andrew and his younger brother Tristan Tate (Pearson 2024b). This paper focuses specifically on the influence of the Tates, and a video titled Emergency Meeting 63: Collapse, posted by the brothers to video-hosting website Rumble on 31 July 2024, just two days after the Southport attack. Using a critical discourse analysis developed by Ashley Mattheis (see Mattheis 2018), the paper evidences three key discursive strands in the Tates' post: Strongmen are Needed for the Coming Collapse, Masculinity Has Fallen: Women are to Blame, Multiracialism Means Ruin. The paper argues the Tates sought to escalate digital hate speech into real-world hate crime through three interlinked mechanisms: (1) the normalisation of violent imaginaries, whereby hateful rhetoric shifts the Overton window to make extremist violence thinkable; (2) emotional scripting, in which affective appeals to shame, fear, and resentment encourage men to reclaim status through violence; and (3) the creation of symbolic events, where the Southport attack and subsequent riots were re-framed as evidence of systemic collapse, demanding violent action.

The paper consists of three sections. Section One considers the literature surrounding Tate, and discussion of the term 'extreme misogyny', before a methodology. Section Two engages a discursive analysis of Tate's key narratives and their genealogy. Section Three concludes that while misogyny familiar to manosphere movements online is inherent to the video, Tate is highly tactical: he utilises far right, accelerationist and populist discourse to ensure his messaging resonates widely, while also exploiting his identity as a mixed-race, self-identified Muslim man. However, the set of discursive strands Tate employs is ultimately lacking a core composite, and ideologically incoherent. This implies Tate's strategic aims are not ideological, but material: to recruit youth to the various Tate products promising masculine strength and status. Nonetheless, Tate's extensive use of far-right narratives, even if ultimately cynical, means his content itself is far-right, and Tate is an extremist. The absence of the category of 'extreme misogyny' does not mean Tate is 'only' misogynist, and not extreme. The Tates' rhetoric embodies both extremism and misogyny. The paper refers to Andrew Tate as 'Tate', and primarily references him as the video author. However, it should be noted that Tate's brother Tristan has an important supporting role, and ensures the rhetoric resonates with an explicitly Christian audience base; Tristan says he is Christian, while Andrew Tate has converted to Islam.

Section 1: Extreme Misogyny, Andrew Tate and the Far-Right Riots

In the wake of the riots following the Southport murders, the UK Government suggested expansion of the definition of extremism to encompass 'extreme misogyny' (Catt 2024). This was in part a response to the growing concerns about violence against women (VAW), but also a recognition of the role of gender in the riots themselves. Scrinzi (2024) has drawn attention to the centrality of narratives on white women's 'security' in the contemporary far right, which she terms the 'racialization of sexism'. Such racialised messaging, positioning men of colour, or of immigrant background as a sexual threat are common to white supremacist groups, dating to the Ku Klux Klan (Patton and Snyder-Yuly 2007). In the UK, for instance, the anti-Islam far right mobilised around the threat of so-called 'Muslim grooming gangs' (Cockbain 2013). The Southport riots employed a similar narrative, the murder of children by an alleged asylum seeker functioning as racialised legitimisation for men's hypermasculinity and violence against other asylum seekers (Pearson 2024b). However, the UK Government later dropped the idea for an explicit category of 'extreme misogyny'.

For some years however, both academics and policy makers have highlighted growing misogyny in online communities and platforms, and the role of misogyny in extremism. Kate Manne (2019) has defined misogyny as a practice of patriarchy, less about hatred of women, but about ensuring men's power and control. Scholars have pointed to the prevalence of misogynistic men's groups, particularly online, through the so-called 'manosphere', misogynist incels or groups such as

Men Going Their Own Way (MGTOW) (Perliger et al. 2023; Kaiser 2022; Barrat 2025; Rothermel 2023; Ging 2019). In a UK context, senior police chiefs have linked online spaces to what they label an 'epidemic' of misogynist violence against women (NPCC 2024a). Extremism meanwhile is variously defined (Schmid 2013; 2023; Commission for Countering Extremism 2019); yet one of the most persuasive definitions of recent years comes from US scholar JM Berger, who suggests there is an objective quality to extremism beyond its relationship to 'mainstream' norms. For Berger (2018), extremism is about an in-group's definition of success residing in hostile actions towards an out-group. Across both the far right and jihadist groups, feminist scholars have noted the importance of misogyny as intrinsic to the: cultures, ideology, organisational objectives and leadership of violent extremist groups (Roose and Cook 2022; Díaz and Valji 2019), as a gateway to extremist activity (United Nations 2021; Phelan et al. 2025); and as intrinsic to the 'hostile action' Berger notes constitute extremism (Pearson 2023).

Contemporary ideological threats globally are primarily the far right and jihadism (Institute for Economics and Peace 2024). While Tate himself is a concern to police in English speaking countries (NPCC 2024b), teachers (Dimsdale 2023; Haslop et al. 2024), academics and women (Leeming 2023; Popa-Wyatt, n.d.), due to his misogynist and abusive language concerning women (Weale 2023; Regehr 2022), he is rarely regarded as either an 'extremist' or as 'political'. In a UK context, this appears at least in part because the focus of the extremism definition has, since 2024, been 'ideology' (UK Government 2024). Specifically, the 2024 definition notes, "(E)xtremism is the promotion or advancement of an ideology[footnote 3] based on violence, hatred or intolerance". However, misogyny and patriarchy are rarely understood in mainstream policy or academia as ideological. Indeed Hoffman, Ware and Shapiro's (2020) assessment of misogynist incel violence noted "the incel worldview is not obviously political" (p.565), suggesting it was closer to "hate crime". A UK Commission for Countering Extremism (CCE) report on incels meanwhile emphasised the mental health needs of this community and suggested they did not broadly support violence, or necessarily believe in the 80:20 rule, a core tenet of incel thinking suggesting the 'top' 80% of women are attracted to the 'top' 20% of men (Whittaker et al. 2023). Nonetheless, Perliger, Stevens and Leidig (2023) have made the case for understanding 'extreme misogyny' as a property of diverse ideological actors who "legitimise violence and measures of coercion against women, and manifest an intense hostility towards symbols of women's empowerment and equality, feminist institutions, and other social constructs that its members feel are threatening to masculinity" (p.11). At the same time, Brace, Baele and Ging (2024) have found manosphere actors online active in disseminating links to material outside their core ideology. This is a phenomenon the authors suggest is relevant to the contemporary trend for Mixed, Unclear and Unstable (MUU) ideologies.

The Southport case is an example in which diverse online actors, including the misogynist influencers the Tates, mobilised to support or encourage violence.

While anti-Islam activist Tommy Robinson called for British men to train for 'resistance' following Southport, and 'X' owner Elon Musk framed Southport as a sign of government failure and the riots similarly as a symbol of resistance, the Tate brothers addressed the riots as a symptom of 'collapse'. They posted Emergency Meeting 63: Collapse to 'TateSpeech' on video-hosting site Rumble on 31 July 2024. 'TateSpeech' was founded in August 2022 and has 2.2 million followers. Collapse is two hours four minutes and 39 seconds long, including some 32 minutes of count-down into the content itself; it has over 10,000 likes, over a thousand comments, and has been viewed nearly a million times, making it one of Tate's most popular broadcasts. This paper is concerned with the nature and role of the rhetoric employed in Collapse and how it relates to wider discourses.

Methodology

The paper now moves to an analysis of the Collapse video, and the role gender plays in the discursive 'scripts' it employs, and the origins of these discourses. Bouvier and Machin (2018) note that "Discourses provide the 'scripts' for acting in society, and in turn social practices embody discourse in the material world that we meet". Specifically, the paper sought to answer the following questions: what gendered narratives does Tate employ? What discourses do they draw on? What is the discursive composite linking the narratives and persuading the audience? What is Tate's central ideological argument? The methodology employed here is drawn from the work of gender, extremism and communications scholar Ashley Mattheis. Mattheis' (2018) paper on far-right female influencer Lana Lokteff explored 'discursive composites', identifying "the multiple strands of discourse synthesized into a whole within the ideological claims forwarded to recruit individuals" (p.135). The aim is to situate discursive strands in their wider cultural narratives.

This method allows for a positioning of Tate's ideas and narratives in wider context. This leads to an argument that Tate's work is clearly ideological, however, that Tate himself draws from disparate and inconsistent discourses, apparently in order to maximise his reach and recruitment. Tate himself exploits ideology, including far right extremist ideology, apparently first and foremost for material gain. Whatever his motives, and in spite of his identity as a mixed-race Muslim (he asserts), his engagement with far-right ideology and narratives allows for a categorisation of 'far right'. However, Tate does not amend his position on gender - which is an assertion of contemporary women's fundamental worthlessness - to accommodate far-right ideology. While subjugating women, traditional far-right narratives also allot them an important maternal role, requiring men's protection.

Section 2: Cover your Bases: The Rumble Recruitment Strategy

This section illustrates how Tate's discursive strands — Strongmen are Needed for the Coming Collapse, Masculinity Has Fallen: Women are to Blame and Multiracialism Means Ruin —draw on far-right, manosphere, and accelerationist discourses. By presenting violence as both inevitable and righteous, and by invoking fear, shame, and emasculation, Tate draws on a diverse, yet incoherent, set of narratives to resonate as widely as possible with male youth and translate into mobilisation. This is the template for action of a manfluencer for whom ideological resonance is instrumentalised primarily for commercial gain. This lack of ideological coherence permits Tate to use extremist rhetoric appealing to far-right, but also jihadist sympathising young men, positioning men as 'warrior protectors', yet lacking women decent enough for men to protect. In Tate's figuration, men are inspired to violence not to protect women of value, but in spite of women who have no value; this is an inversion of jihadist and far right traditional gender norms. The findings suggest Tate seeks to escalate digital hate speech into real-world hate crime through three interlinked mechanisms: (1) the normalisation of violent imaginaries, whereby hateful rhetoric shifts the Overton window to make extremist violence thinkable; (2) emotional scripting, in which affective appeals to shame, fear, and resentment encourage men to reclaim status through violence; and (3) the creation of symbolic events, where attacks or crises are re-framed as evidence of systemic collapse, demanding action. This is achieved through three discursive strands: 1) Strongmen are Needed for the Coming Collapse 2) Masculinity Has Fallen: Women are to Blame 3) Multiracialism Means Ruin. Timestamps for speech understand 0'0 as the start of the speech section of the video. It should also be noted that the analysis reproduces transcript from Collapse, which is offensive.

Strand One: Strongmen are Needed for the Coming Collapse

The video's first discursive strand frames England's far right riots in the wake of the Southport attacks as evidence of the inevitability of coming western democratic 'collapse', and to articulate the only correct role for men and manhood: physical strength and (hyper)masculinity expressed in domination and violence. Collapse is articulated as an outcome of western decadence; for Tate the viewer should understand that institutions cannot be trusted and: "judges are not fair and that courts are not fair, and that the law is not real, and that elections are not real, and your vote means nothing more, and that the banking system is not real. Democracy collapses under this knowledge" (8'35-8'47). This message is one Tate has often repeated in reference to his own situation, in which he claims to be a victim of the so-called 'Matrix', seeking to falsely pursue him for spurious charges, including sexual assault, trafficking and rape. Democracy is a façade, and only one thing is real, men's ability to fight: "all that's left behind democracy in the first place, behind all these fake institutions, is violence, is force" (8'52-7). This force is necessarily hypermasculine, "The change in culture and consciousness which is required is men and masculinity and men of honour. ... you need men who are prepared to be men, because men are the protectors of society" (20'-20'29). When Collapse comes, Tate believes this will be both racialised and gendered, because democracy has been weakened through multiculturalism and feminism:

> So it's going to turn to violence. The Western world is going to get very violent. As more and more people wake up to the idea and the understanding that the whole democratic process and the democratic ideal itself is corrupt. What was it Enoch Powell said - the rivers will run red with blood, and everyone called him a racist…. we gave power to a bunch of fucking clowns and women and they fucked it up. (16'47-17'12).

Ideologically, Tate owes rhetoric to British MP Enoch Powell's 1968 speech, referenced here, protesting immigration and the introduction of the Race Relations Act. While denounced by his peers as racist, and leading to his sacking, this speech won Powell some popular support from working-class men (Whipple (2009) cited in Crines et al. 2016): Tate's 'common man'. Tate also invokes far-right accelerationism, which asserts that liberal institutions are weak and collapse must be hastened to bring about a new world order (Jipson 2025). As Jipson (2025) notes, accelerationism requires 'social chaos' which "creates an opportunity for extremists to create a racially or ideologically "pure" future". As such the narrative is linked to the 'Great Replacement' conspiracy's assertion that white people are victim to a strategy to see them lose power to non-white people through replacement: in immigration and rising non-white birth-rates. Tate's messaging additionally echoes far- and radical-right populist messaging, asserting that elites are corrupt and cannot be trusted (Krzyżanowski and Ekström 2022; Meret and Siim 2013). These narratives have been employed for instance in President Trump's election campaign, or in Reform Party rhetoric in the UK, and are mainstream.

Tate references the Strong-man Weak-man cycle, in which a period of 'weak' men, feminised through persistent peace, necessitates a period of 'strong' men and war. The idea of success through masculine strength has historical roots dating back centuries (Devereaux 2025). It has been revived in recent authoritarian discourse thanks to a post-apocalyptic 2016 novel by G. Michael Hopf (2016), containing the lines "Hard times create strong men, strong men create good times, good times create weak men, and weak men create hard times" . While Tate echoes far right messaging positioning men's racialised violence as valour, he suggests this struggle is for society, and other men; it is not for wives, Elshtain's (1987) 'beautiful souls' that 'just warriors' must fight to protect, or for family. Tate explicitly exploits post-Southport violence as a symbol of democratic breakdown equivalent to Hopf's apocalypse, and the inevitable anger of the 'common man'. This emotion is proposed as a powerful resource towards a hypermasculinity widely understood as hegemonic: the violent strongman. As such Tate taps into discourses that straddle mainstream populist narratives, as well as culturally resonant ideas of men's powers, and the accelerationist far right, all of which are highly prevalent online and will be familiar to his audience.

Strand Two: Masculinity Has Fallen: Women are to Blame

If men have been weak, and mobilised into action only when the system is collapsing, as evident in the violence in Southport, the responsibility is not theirs alone. The second discursive strand addresses the causes of men's current weakness. While the far-right has often mobilised riots around the gendered imperative of substituting for ineffective states, in order to counter 'immigrant' violence against 'native' women and children (Scrinzi 2024), Tate centres the inability of ordinary men to perform manhood in women's failure, not men's. He says:

> You don't need people like me to defend the population and pretend to protect the society. You need the common man. But the problem is, you women have become so awful that the common man has no interest at all in protecting you, because he no longer has a family, no longer has a wife that obeys. (18'20-28)

This rhetoric has a powerfully broad recruitment base, given its basis in both Islamic & Christian traditionalism, representing the faiths each Tate brother says he follows. The narrative suggests the importance of the gender binary and complementarity of men's and women's roles is not just natural, but God-given, and focused on obedience. The appeal is therefore to the global populations of young men who believe in the enduring and God-given unequal roles for men and women within relationships. The appeal however is not to the anti-Islam radical right, who espouse gender equality as a means of demonstrating the backwardness of Islam, and whose key focus is the sexually predatory Muslim man (Cockbain 2013; Pilkington 2016); it is to more fundamentalist proponents of right wing ideology, and the online manosphere.

Women's failure to perform the binary role of traditional nationalism is centred for Tate on their lack of respect for men's superiority as 'head of the house', and by implication their agency, enabled through feminism, a central theme in his messaging. If valour is absent in the 'common man', it is not because the warriors are not just, but because the souls they have historically protected are no longer beautiful. This rhetoric is additionally a means of distinguishing the brothers –who Tate positions as an already strong elite - from the majority of 'common' men, who are unable to perform strength and masculinity. Tate explicitly links [common] men's failure to protect, to women's contempt for the common man [which Tate is not]. The 'common man': "*can't find a wife who obeys him. And if he does get a wife, she's such a nasty person to him all the time that he doesn't really care about it. …. Why would he protect people who despise him?*" (18'40-19'14) Women's disobedience is however merely symbolic of wider societal disdain for the 'common man', which Tate contrasts with past periods of war: "*That's what men died for. They went to war for their women…(19'29) ..Your common man has checked out. And that's because women have become so ridiculous and lost their minds with unfair expectations…*" (19'55-20'03).

Women's infidelity, their claimed 'ridiculousness', is both sexualised and racialised. Tate notes that women's sexuality is often – perfidiously - focused on men of colour, therefore transgressing the binary of racialised nationalism: white men fight to protect women from the hypermasculinity of immigrant non-white men, from sexual violence and from miscegenation. If native women instead pursue immigrant men, the gendered binary logic of nation is inverted, Collapse is inevitable. Specifically post-Southport, white English 'native' men no longer have either status, or a role. Women's faithfulness is linked discursively to social modernity; their faithlessness is a key cause of Collapse. Use of the word 'treason' emphasises the asserted linkage between women's lack of obedience, and western society's fall. Tate says:

> Sexual access is the primary motivator for all things for men. …It's why they built the whole modern world …. Effectively, to have unlimited sexual access. So once that's cut off from them, they have no point in doing many things at all, especially not risking their freedom and safety to protect a society which has women now believing that they're superior to 98.5% of men, insulting 98.5% of men, then complaining those men don't stand up, defend them when the invaders turn up. On top of that, what's most treasonous is that a lot of these women are perfectly happy to embrace the invaders. (20'28-20'56)

This section of the video directly references misogynist incel ideology in which women are understood primarily in terms of 'sexual access'. Similarly, in Tate's view, women's core worth, not just to men, but to wider society, is in subservience, specifically sexual subservience. Such language is therefore a direct appeal to, and repetition of, misogynist incel community rhetoric, asserting core incel ideology centred on the belief that women are sexually attracted to only a small percentage of high-status, physically powerful men, leaving weak men involuntarily celibate (Leidig 2021; Ging 2019).

Through this narrative Tate asserts the modern feminised society as not just weak but racially impure. Southport and the subsequent violence are constructed as a symbol of a gendered breakdown. This is not a 'crisis of masculinity' - as none of this is men's fault - but of women's failed fidelity, which results in the sexual frustration and emasculation of the common man, who must cede sexual status and strength to the racialised invader. This is a common recruitment tactic for neo-Nazi and white supremacist far right groups (Kimmel and Ferber 2000), and would be recognised as such by these communities. The strategy here is to denigrate the common man, while shifting responsibility for this denigration from Tate himself to apparently faithless women. Tate's video narrative effectively employs emotional scripts designed first to stoke men's shame, fear, and resentment, and then to make affective appeals to these emotions, encouraging men to reclaim status through violence (Bengtsson 2016).

Strand Three: Multiracialism Means Ruin

The third discursive strand emphasises the racialised, as well as feminised, nature of the weak, collapsing society that has been a key argument of the second strand. Now, the Tate brothers use this argument to assert their own superiority over the common man, and his unwillingness to fight. "Wake the fuck up", Tate demands. "*The average Englishman will go to fight over his football team, but* *won't fight to defend his fucking society. It's bullshit. … It's over. Unless you stand up. .. the decimation of Christianity and the decimation of the Western world in real time, that's all I see*" (12'56-13'22). Tate frames this as a "harsh truth" (13'22), establishing his role as mentor, advisor and the best friend who does not mince his words, a key function of his wider messaging, and the core product he sells to young men: access to networks of mentors in business (Pearson 2024a).

Referencing Southport Tristan develops the narrative, noting the attacker was not, as first thought, Muslim, (misinformation in fact shared online by Tate, amongst others). Tristan shifts the rhetoric away from faith towards race, equating black masculinity with 'savagery' 'capacity for violence' and physical superiority, while also emphasising the black man as a 'Third Worlder', lacking 'modernity' and the civilisational qualities that the rest of the video has been at pains to urge the common man to protect:

> The whole anti-Islamic rhetoric of 'Muslims are dangerous, Muslims are savages', etc., etc. actually was proven not to be so true in the recent terrorist attack, and what me and you always say has been proven true, which is - people from the Third World are savages! (15'36-52)…. Everyone needs to keep the Third Worlders out of the First world. Every country! Because this guy who stabbed all those kids was from Rwanda! (16'30-48)

Earlier, Andrew Tate had invoked insider knowledge to highlight the capacity of Black people for violence: "T*hese people are built different. They smoke tear gas. I'm telling you, as someone who's mixed-race, Black people are built different. We have a different capacity for violence. You're not ready for us. You need to fucking train up if you want to fight us hand-to-hand*" (8'06-16). The brothers invoke white supremacist narratives of racial difference, promoting fear of the black man's hypermasculinity, and hypersexuality (Kimmel 2017). However, these sections of Collapse exploit Southport to also assert the superiority of the Tates, as black men, through narratives of black men's capacity for violence. *Collapse* reveals the brothers' negotiation of their own mixed-race heritage and identity. As Joseph Salisbury (2019) has observed in work on Black men in the US and UK – the Tates are nationals of both - Black mixed-race men must contend with the 'white gaze', which casts black masculinity as a 'monster', the black man: "criminal, violent, unintelligent and a sexual predator" (p.1770). While at times they contest the racist projections of the white gaze, Joseph Salisbury also notes, "In other instances, Black mixed-race men may choose to utilize that stereotype, whether to gain popularity or as protection from racism and bullying" (p.1770). The popularity of the *Collapse* video, and the Tates online is important as a means to sell products including their online training network 'The Real World', which the brothers devote the latter portion of the broadcast to promoting.

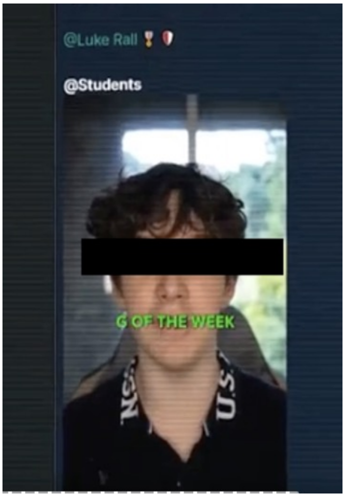

[redacted]

Figure 1. G of the Week

While the Tates' psychological motivations are unclear, the genealogy of such narratives is evident. The brothers reference a range of white supremacist tropes, including Accelerationism, and the Great Replacement Theory, as explored in the previous discursive strand: cultural and racial nationalism, in which 'the First World' is under threat; racial and biological essentialism, positioning black masculinity as monstrous and exhibiting a surplus of strength and sexuality. The strategy is also made clear at the end of *Collapse*, when the viewer is invited to buy products. The denigration of the common man as weak, needing to be mentored and trained by the Tates, who have insider knowledge of the 'Black male invader' pushes the viewer to invest in what the Tates are selling. The brothers exploit their mixed-race status, engaging in racial essentialism for commercial gain. While all of the video to this point has discussed 'men', the final sections – including that in which star student/customer 'G OF THE WEEK' is depicted, reveal that the true audience is in fact not men at all, but youths and boys (see Figure 1).

Conclusion: Tate as Discursive Composite

The paper has used a critical discourse analysis to evidence how the Tates embed recruitment narratives in widely resonant discourses. They exploit white supremacist, manosphere, incel, far right, religious hegemonic Judeo-Christian as well as populist gendered discourses, framed to appeal to the widest male audience. They effect is to synthesise these mixed discourses into a unifying affective theme in which Tate himself represents the composite of three discursive strands employed in the *Collapse* video: *Strongmen are Needed for the Coming Collapse, Masculinity Has Fallen: Women are to Blame and Multiracialism Means Collapse*. These discursive strands link the need for strong men to a failed state, corrupted through racialisation and feminisation. In this state, men no longer have worth; they must fight to reclaim this, becoming 'strong men'. The Tates are primarily engaged in the normalisation of violent imaginaries, whereby hateful rhetoric shifts the Overton window to make extremist violence thinkable; they do this by invoking familiar scripts from far right and white supremacist rhetoric, without explicitly casting themselves as adherents of far-right ideology.

In fact, the Tates navigate the tension in their use of white supremacist narratives, as Black mixed-race men, in two ways: first, by failing to adapt their gender stance to the traditional far-right. For the Tates, men are inspired to violence not to protect women of value, but in spite of women who have no value. Second, by simultaneously aligning their own identities with the physically superior hypermasculinity of the Black 'Third Worlder'. *Collapse* seeks to mobilise youth and men through emotional scripting and the Tates' emotional dominance, in which affective appeals to shame, fear – of the racialised other but also of their own weakness - and resentment encourage men to reclaim status through violence. Third, the Tates construct the Southport attack as a symbolic event, re-framing this as evidence of systemic collapse, demanding men's action. The men also leverage their own proclaimed religious faith to further legitimise their vision of resistance, noting, "*The resistance are not the Twitter accounts talking about imaginary deportations, the resistance are never going to be the mainstream people because they've sold their souls. The resistance is going to be you, you taking action, which is what God wants you to do*" (16'54-17'06). The necessary action is to subscribe to the Tates' online products which promise brotherhood, money, 'hot bitches', cars and power.

The solution the Tates offer is the image of the 'Strongman'. While the Tates exploit affective arguments of cowardice, fear, shame, lack of virility to prompt action, the ultimate authority and legitimisation they offer is religious. In the latter stages of the video, Tate shows footage of Christian strongmen 'powerlifters' performing to rock music (see Figure. 2), TV the brothers apparently enjoyed as children in the United States. Throughout Collapse the Tates cherry pick from popular online extreme ideologies engaging misogyny in order to maximise resonance across - youth - audiences.

Figure 2. Tate plays 'Christian Powerlifter' videos from his childhood

Scholars and policy-makers primarily engage and identify Andrew and Tristan Tate as 'misogynists', the household names of the manosphere, who put the manosphere on the agenda of schools, universities and parents (Weale 2023; Wescott et al. 2024). This paper however emphasises that while the Tates are clearly misogynist and anti-feminist - they promote patriarchy, male supremacy and the denigration of women's agency - they are also 'extremists' in the broader sense of the term. The Tates may not, as Black mixed-race men, believe in white supremacist or far right ideology, but they are certainly open to propagating this for their own gain. As recognition of Mixed Unclear Unstable or Complex ideologies in recent years has demonstrated, a coherent ideology is not necessary to being 'extreme' (Brace et al. 2024). The Tates' role in stoking anti-democratic unrest, far right politics and violence for their own commercial gain deserves as much recognition as their misogyny and gender-based violence, online and off.

References

Barrat, Erin. 2025. 'Tackling Extremist Misogyny in the Digital Age'. *Humanities News.* University of Manchester, September 26. https://www.manchester.ac.uk/about/news/tackling-extremist-misogyny-in-the-digital-age/.

Bengtsson, Tea Torbenfeldt. 2016. 'Performing Hypermasculinity: Experiences with Confined Young Offenders'. *Men and Masculinities* 19 (4): 410–28. https://doi.org/10.1177/1097184X15595083.

Berger, J. M. 2018. *Extremism*. MIT Press.

Brace, Lewys, Stephane J. Baele, and Debbie Ging. 2024. 'Where Do 'mixed, Unclear, and Unstable' Ideologies Come from? A Data-Driven Answer Centred on the Incelosphere'. *Journal of Policing, Intelligence and Counter Terrorism* 19 (2): 103–24. https://doi.org/10.1080/18335330.2023.2226667.

Catt, Helen. 2024. 'Misogyny to Be Treated as Extremism by UK Government'. *BBC News*, August 18. https://www.bbc.co.uk/news/articles/c15gn0lq7p5o.

Cockbain, E. 2013. 'Grooming and the "Asian Sex Gang Predator": The Construction of a Racial Crime Threat'. *Race & Class* 54 (4): 22–32. https://doi.org/10.1177/0306396813475983.

Commission for Countering Extremism. 2019. *Challenging Hateful Extremism*. CCE.

Crines, Andrew, Tim Heppell, and Michael Hill. 2016. 'Enoch Powell's "Rivers of Blood" Speech: A Rhetorical Political Analysis'. *British Politics* 11 (1): 72–94. https://doi.org/10.1057/bp.2015.13.

Devereaux, Bret. 2025. 'Hard Times Don't Make Strong Soldiers'. *Foreign Policy,* November 25. https://foreignpolicy.com/2020/05/02/hard-times-dont-make-strong-soldiers-warrior-myth/.

Díaz, Pablo Castillo, and Nahla Valji. 2019. 'Symbiosis of Misogyny and Violent Extremism: New Understandings and Policy Implications'. *Journal of International Affairs* 72 (2): 37–56.

Dimsdale, Connie. 2023. 'How Andrew Tate Is Still Influencing Boys in Schools Who Believe His Arrest Is a "Conspiracy"'. *Inews.Co.Uk*, March 19. https://inews.co.uk/news/andrew-tate-influencing-boys-schools-arrest-conspiracy-2213665.

Dodd, Vikram, Ben Quinn, and Rowena Mason. 2024. 'Former Counter-Terror Chief Accuses Farage of Inciting Southport Violence'. UK News. *The Guardian*, July 31. https://www.theguardian.com/uk-news/article/2024/jul/31/farage-accused-of-inciting-southport-violence-by-former-counter-terror-chief.

Elshtain, Jean Bethke. 1987. *Women and War*. University of Chicago Press.

Ging, Debbie. 2019. 'Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere'. *Men and Masculinities* 22 (4): 638–57. https://doi.org/10.1177/1097184X17706401.

Haslop, Craig, Jessica Ringrose, Idil Cambazoglu, and Betsy Milne. 2024. 'Mainstreaming the Manosphere's Misogyny Through Affective Homosocial Currencies: Exploring How Teen Boys Navigate the Andrew Tate Effect'. *Social Media + Society* 10 (1): 20563051241228811. https://doi.org/10.1177/20563051241228811.

Hoffman, Bruce, Jacob Ware, and Ezra Shapiro. 2020. 'Assessing the Threat of Incel Violence'. *Studies in Conflict & Terrorism* 43 (7): 565–87. https://doi.org/10.1080/1057610X.2020.1751459.

Hopf, G. Michael. 2016. *Those Who Remain: A Postapocalyptic Novel:* Volume 7 (The New World Series). CreateSpace Independent Publishing Platform. https://www.amazon.co.uk/Those-Who-Remain-Postapocalyptic-Novel/dp/1539031314/ref=monarch_sidesheet_title.

Institute for Economics and Peace. 2024. *Global Terrorism Index 2024 - World* | ReliefWeb. Institute for Economics & Peace. https://reliefweb.int/report/world/global-terrorism-index-2024.

ISD. 2024. 'From Rumours to Riots: How Online Misinformation Fuelled Violence in the Aftermath of the Southport Attack'. *Digital Dispatches*, July 31. https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/.

Jipson, Art. 2025. 'A Field Guide to "Accelerationism": White Supremacist Groups Using Violence to Spur Race War and Create Social Chaos'. *The Conversation*, June 11. https://doi.org/10.64628/AAI.5scdetctw.

Joseph-Salisbury, Remi. 2019. 'Wrangling with the Black Monster: Young Black

Mixed-Race Men and Masculinities'. *The British Journal of Sociology* 70 (5): 1754–73. https://doi.org/10.1111/1468-4446.12670.

Kaiser, Susanne. 2022. *Political Masculinity: How Incels, Fundamentalists and Authoritarians Mobilise for Patriarchy*. 1st edition. Translated by Valentine A. Pakis. Polity.

Kimmel, Michael. 2017. *Angry White Men: American Masculinity at the End of an Era*. Revised edition. Bold Type Books.

Kimmel, Michael, and Abby L Ferber. 2000. 'White Men Are This Nation': Right-Wing Militias and the Restoration of Rural American Masculinity. 65 (4): 582–604.

Krzyżanowski, Michał, and Mats Ekström. 2022. 'The Normalization of Far-Right Populism and Nativist Authoritarianism: Discursive Practices in Media, Journalism and the Wider Public Sphere/s'. *Discourse & Society* 33 (6): 719–29. https://doi.org/10.1177/09579265221095406.

Leeming, Megan Cait. 2023. *Radicalised Masculinity: Ontological Insecurity, Extremist Ideologies and the Rise of Andrew Tate.* September 21. https://dspace.cuni.cz/handle/20.500.11956/187377.

Leidig, Eviane. 2021. *Why Terrorism Studies Miss the Mark When It Comes To Incels*. August 31. https://icct.nl/publication/why-terrorism-studies-miss-the-mark-when-it-comes-to-incels/.

Manne, Kate. 2019. *Down Girl.*

Mattheis, Ashley. 2018. 'Shieldmaidens of Whiteness: (Alt) Maternalism and Women Recruiting for the Far/Alt-Right'. *Journal for Deradicalization* 0 (17): 128–62.

Meret, Susi, and Birte Siim. 2013. 'Gender, Populism and Politics of Belonging: Discourses of Right-Wing Populist Parties in Denmark, Norway and Austria'. In *Negotiating Gender and Diversity in an Emergent European Public Sphere*, edited by Birte Siim and Monika Mokre. Gender and Politics Series. Palgrave Macmillan UK. https://doi.org/10.1057/9781137291295_5.

NPCC. 2024a. 'Call to Action as VAWG Epidemic Deepens'. *News Releases.* National Police Chiefs' Council (NPCC), July. https://news.npcc.police.uk/releases/call-to-action-as-violence-against-women-and-girls-epidemic-deepens-1.

NPCC. 2024b. Violence Against Women and Girls (VAWG) National Policing Statement 2024. *National Policing Statement. NPCC.* https://cdn.prgloo.com/media/5fc31202dd7e411ba40d29fdca7836fd.pdf.

Patton, Tracey Owens, and Julie Snyder-Yuly. 2007. 'Any Four Black Men Will Do: Rape, Race, and the Ultimate Scapegoat'. *Journal of Black Studies* 37 (6):

859–95. https://doi.org/10.1177/0021934706296025.

Pearson, Elizabeth. 2023. *Extreme Britain: Gender Masculinity and Radicalisation.* Hurst.

Pearson, Elizabeth. 2024a. 'Misogyny, Misandry and (Online Cult) Leader: The Daily Emails of Andrew Tate'. *VOX* - Pol, July 24. https://voxpol.eu/misogyny-misandry-and-online-cult-leader-the-daily-emails-of-andrew-tate/.

Pearson, Elizabeth. 2024b. 'The Hypermasculine Far Right: How White Nationalists Tell Themselves They Are "Protecting" Women and Children When They Riot'. *The Conversation*, August 7. https://theconversation.com/the-hypermasculine-far-right-how-white-nationalists-tell-themselves-they-are-protecting-women-and-children-when-they-riot-236250.

Perliger, Arie, Catherine Stevens, and Eviane Leidig. 2023. *Conceptualising Extreme Misogyny. Mapping the Ideological Landscape of Extreme Misogyny*. International Centre for Counter-Terrorism.

Phelan, Alexandra, Jessica White, Claudia Wallner, and James Paterson. 2025. 'Gendered Narratives and Misogyny as Motivators Towards Violent Extremism: The Case of Far-Right Extremism in the UK and Australia'. *Terrorism and Political Violence* 37 (7): 961–78. https://doi.org/10.1080/09546553.2025.2498505.

Pilkington, Hilary. 2016. *Loud and Proud: Passion and Politics in the English Defence League*. Manchester University Press.

Popa-Wyatt, Mihaela. n.d. '*Written Evidence Submitted by Dr Mihaela Popa-Wyatt (COM0036)*'. Manchester University. Accessed 16 November 2025. https://committees.parliament.uk/writtenevidence/142834/pdf/.

Regehr, Kaitlyn. 2022. 'In(Cel)Doctrination: How Technologically Facilitated Misogyny Moves Violence off Screens and on to Streets'. *New Media & Society* 24 (1): 138–55. https://doi.org/10.1177/1461444820959019.

Roose, Joshua M., and Joana Cook. 2022. 'Supreme Men, Subjected Women: Gender Inequality and Violence in Jihadist, Far Right and Male Supremacist Ideologies'. *Studies in Conflict & Terrorism* 0 (0): 1–29. https://doi.org/10.1080/1057610X.2022.2104681.

Rothermel, Ann-Kathrin. 2023. 'The Role of Evidence-Based Misogyny in Antifeminist Online Communities of the "Manosphere"'. *Big Data & Society* 10 (1): 20539517221145671. https://doi.org/10.1177/20539517221145671.

Schmid. 2013. Radicalisation, *De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review*. International Centre for Counter-Terrorism, The Hague. http://www.icct.nl/download/file/ICCT-Schmid-Radicalisation-De-Radicalisation-Counter-Radicalisation-March-2013.pdf.

Schmid, Alex P. 2023. *Defining Terrorism*. ICCT Report. International Centre for Counter-Terrorism. https://www.icct.nl/sites/default/files/2023-03/Schmidt%20-%20Defining%20Terrorism_1.pdf.

Scrinzi, Francesca. 2024. *The Racialization of Sexism: Men, Women and Gender in the Populist Radical Righ*t. Routledge.

The Crown Prosecution Service. 2025. '*Teenager Jailed for Killing Three Children at a Dance Class and Trying to Kill Ten Other*s'. https://www.cps.gov.uk/mersey-cheshire/news/teenager-jailed-killing-three-children-dance-class-and-trying-kill-ten.

UK Government. 2024. 'New Definition of Extremism (2024)'. *GOV.UK*, March. https://www.gov.uk/government/publications/new-definition-of-extremism-2024/new-definition-of-extremism-2024.

United Nations. 2021. *Misogny: The Extremistk Gateway? Issue Brief 2. Extremism in Focus.* United Nations Development Programme, Oslo Governance Centre. https://www.undp.org/sites/g/files/zskgke326/files/migration/oslo_governance_centre/Misogyny-The-Extremist-Gateway.pdf.

Watling, Tom. 2024. 'Fact Checked: The False Far-Right Claims That Sparked the Riots'. *The Independent*, August 3. https://www.independent.co.uk/news/uk/crime/southport-stabbing-latest-far-right-riots-b2590454.html.

Weale, Sally. 2023. '"We See Misogyny Every Day": How Andrew Tate's Twisted Ideology Infiltrated British Schools'. Society. *The Guardian*, February 2. https://www.theguardian.com/society/2023/feb/02/andrew-tate-twisted-ideology-infiltrated-british-schools.

Wescott, Stephanie, Steven Roberts, and Xuenan Zhao. 2024. 'The Problem of Anti-Feminist "Manfluencer" Andrew Tate in Australian Schools: Women Teachers' Experiences of Resurgent Male Supremacy'. *Gender and Education* 36 (2): 167–82. https://doi.org/10.1080/09540253.2023.2292622.

Whittaker, Joe, William Costello, and Andrew Thomas. 2023. P*redicting Harm Among Incels (Involuntary Celibates): The Roles of Mental Health, Ideological Belief and Social Networking*. Commission for Countering Extremism.

Goddesses of Vengeance: How Far-Right Feminist Groups Co-opt Women's Rights to Fuel Polarization and Violence

Gwenaëlle Bauvois, University of Helsinki

Introduction

In recent years, far right movements in Europe have undergone a marked process of feminisation. We can witness a reduction in what is known as the Radical Right Gender Gap, the feminisation of the far-right electorate. For instance, in France, the female vote for far-right parties has risen from 20% in 2019 to 30% in 2024 (Bonnefond, 2024; Politico, 2024; Rey & Gonthier, 2024), indicating a growing adherence by women. Political figures such as Giorgia Meloni, Riikka Purra, Marine Le Pen, and Alice Weidel exemplify this transformation. While earlier scholarship had largely neglected women's roles in the far right (Blee & McGee Deutsch, 2012; Bohoslavsky, 2024), more recent research has highlighted their growing place within this space (Eksi, 2021; Della Sudda, 2022; Almony, 2022). In the context of this paper, the far right refers to a family of movements structured around nativism and exclusion, typically advocating ethnocultural homogeneity, rejecting pluralism, and promoting hierarchical conceptions of social order (Mudde, 2007; Mudde & Kaltwasser, 2017; Rydgren, 2007, 2018; Akkerman, 2015; Farris, 2017; Moffitt, 2016). Within this framework, gender plays a central role, originally dominated by male figures and masculinist rhetoric, far-right movements are now increasingly populated by women who strategically employ the language of women's empowerment. They do so to advance a femonationalist agenda (Farris, 2017)—a political strategy that instrumentalizes gender equality to justify nationalist, anti-immigrant, and particularly anti-Muslim stances.

This general feminisation process within the far right political sphere has been accompanied by the feminisation of female grassroot militantism operating at the close periphery of the political far right. This rise must be understood within what scholars describe as a "reactionary turn" in feminism (Kay, 2024) and a broader crisis of left-wing political discourse (Vendrand-Maillet, 2020). As progressive movements are increasingly perceived as detached from everyday concerns such as security, precarity, and cultural identity, some female far-right actors have succeeded in appropriating progressive rhetoric to turn into reactionary rhetoric and advance exclusionary agendas (Della Suda, 2025). One category of such far-right movements is still largely understudied: grassroot far right feminist movements led by and for women (Calderaro, 2025; Alestra, 2025; Della Sudda, 2022; Debras, 2022; Goetz, 2022).

What does far-right feminism refer to here? While self-identifying as feminist, these groups mobilise a femonationalist logic, framing migrant men as an inherent threat to justify exclusionary politics under the guise of defending women's rights. In doing so, feminism is reframed as a project of civilisational defense rather than social equality. In this article, we focus on the French far-right feminist group called Collectif Némésis as it provides not only a clear illustration of this phenomenon but also underscores the tangible connection between digitally propagated hate speech and real-world actions. Indeed, Némésis and similar groups are particularly active in the digital space (François, 2021) but does not stay confined to digital spaces. Némésis conducts street-level actions—including unauthorised participation in marches, staged performances, and masked gatherings—sometimes in coordination with other far-right groups, some of which have a history of violence. This has already led to multiple legal complaints for incitement to hatred, harassment, and defamation. This creates a pipeline from digital discourse to tangible offline harm and demonstrates the increasing permeability between digital hate and physical action.

This study is guided by the following research question: How do far-right feminist movements' actions demonstrate a tangible connection between online hate speech and offline mobilisation? Specifically, how do their alliances with violent far-right actors and the organisation of physical events translate a digital ideology into concrete action, thereby blurring the line between discourse and potential incitement?

Methodology and data

This study mobilises a multimodal digital ethnographic approach to examine Collectif Némésis as a contemporary expression of far-right feminism operating both within digital and physical environments (Berry, 2012; Jouët and Le Caroff, 2013; Pink et al., 2016; Hine, 2020). As Pink et al. (2016) argue, digital ethnography is not confined to the study of online spaces per se but rather attends to how digital practices are interwoven with everyday life and political subjectivities. This methodological orientation allows for an analysis that captures both the circulation of discourse across platforms and the affective, embodied engagements that sustain offline communities.

Central to this approach is non participant observation (Hine, 2015). By following Collectif Némésis's activities across all their social media accounts: Telegram, Instagram, X, Tik Tok and Facebook, the researcher can observe how narratives, linguistic repertoires, and interactional norms are produced, maintained, and contested. Complementing this, discourse and visual analysis serves to interpret the group's communicative strategies and symbolic repertoire. Drawing on multimodal approaches (Kress & van Leeuwen, 2001), this method involves the examination of memes, slogans, iconography, and hashtags to understand how the movement constructs a coherent ideological and aesthetic identity. The corpus consists of a diverse set of content produced by the Collectif Némésis (mostly in France but also beyond) as an association and by its main representatives individually (2019-2025). It includes texts, photographs and videos disseminated on the following digital platforms: Telegram, X, Instagram, Facebook, TikTok and YouTube. We also collected and analysed interviews given by members of Némésis in podcasts and other audiovisual media (radio and television shows, and 'reinformation' media outlets) as well as appearances in media programmes (radio shows and other media platforms where Némésis members are regular hosts). This corpus is complemented by news articles about Némésis and/or featuring Némésis, in legacy media (from the left to the right), as well as far right reinformation media.

Collectif Némésis: a new feminism?

Reactionary feminism is nothing new. Already in the 70's, a form of conservative feminism already existed, with some women mobilising to defend traditional gender roles and oppose reforms such as abortion liberalisation. Notable examples include Catholic women's movements in France, Italy, and Germany, and the US STOP-ERA campaign led by Phyllis Schlafly. Scholars now analyse these movements as early expressions of conservative and reactionary feminism (Critchlow 2005; Masquelier 2019; Tichenor 2020; Venner, 1993).

Even the phenomenon of far right feminism we are referring to in this paper is not new as such. In the context of France, the real turning point in right and far-right female activism came around 2012 with the rise of La Manif pour Tous, formed in response to the Taubira law legalising same-sex marriage. Many scholars identify La Manif pour Tous as the catalyst for this new wave of right-wing and far-right mobilisation around gender (Raison du Cleuziou, 2019; Della Sudda, 2022). The Belle et Rebelle collective (2011) which promoted pride in European roots while rejecting both leftist feminism and multiculturalism, served as a basis for Némésis. As well as Les Caryatides, founded in 2013 and linked to L'Œuvre Française (a far-right movement in France, dissolved in 2013), which advocated a return to traditional gender values while denouncing the 'denaturalisation' induced by contemporary feminism. It is important to note that this far-right feminist grassroot activism is of course not limited to France, though it seems to be more prominent, examples can be found elsewhere, such as Germany since the 2015 refugee crisis (Camacho, 2021). We can name the Lukreta 'Initiative' (Goetz, 2022; Scheyer et al., 2025) a modern far-right recoding of a mythological symbol of violated feminine purity and national uprising. Similar to Némésis, Lukreta frames immigrants as a threat for women's safety in the public sphere and both groups have collaborated and coordinated public actions together. Ultimately, Nemesis does not represent a radical departure from existing paradigms. However, what is striking is the new visibility in the public debate and mediatisation of the far-right feminism they seem to embody. According to some, this movement is unique in that it is the first far-right women's movement to have been founded by and for women and to have gained momentum (Plottu & Macé, 2024). At the very least, we can agree that Némésis, while not the first of its kind, has achieved unprecedented visibility in the public sphere.

This leads to a central question regarding the Collectif Némésis: what constitutes its central tenets and strategic goals? Founded in 2019 by a small group of young women, Némésis was first aiming at exposing sexual and gender-based violence in the public space. Némésis emerged from discussions on the Facebook page Bellica and the private messaging channel Bellica Paris, both managed by the identitarian influencer Solveig Mineo (Della Sudda, 2025) who called upon a certain form of feminist sisterhood. Central to the group's discourse is a systematic association of immigration, Islam, and sexual violence. Némésis present themselves as «whistleblowers» and as an « association for women's protection » and more recently even as an « association for victims » of sexual violence, while critiquing what it perceives as the complicit silence of institutional feminism regarding violence perpetrated by foreign men. It is difficult to know how large the Collectif Némésis actually is, especially when considering core activists and supporters. Estimates of Némésis's membership vary greatly depending on the source. The President of the Collectif Alice Cordier has stated that the group grew from six active members (core activists) in 2019 to 450 in 2025, while different media sources suggest roughly 150 active members in France, with the operational core likely numbering between 30 and 80 (Msika, 2025; Pernes, 2025). The collective currently employs five staff members and, since summer 2025, has eight official spokespersons.

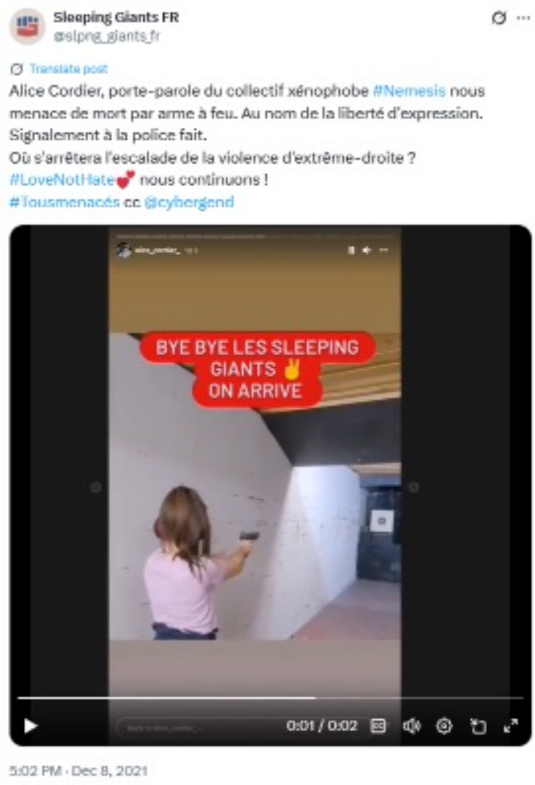
Némésis' significant influence and growth is not incidental but stems from a confluence of strategic factors. Firstly, it benefits from a robust organisational framework and from the patronage of prominent business figures such as Vincent Bolloré and Pierre-Édouard Stérin. Bolloré, a powerful media magnate whose corporate empire encompasses outlets like CNews, Europe 1 and Le Journal du Dimanche which play a central role in amplifying right-wing and reactionary discourse. Stérin, a conservative billionaire known for financially supporting identitarian and traditionalist initiatives. Together, Bolloré and Stérin have even convened political "influence" events—such as a recent 'summit of freedoms'—designed to unite right-wing and far-right actors around a shared conservative and nationalist agenda. Secondly, Némésis' operations are marked by a high degree of professionalisation, particularly in its sophisticated media strategy. This includes a proven capacity for crafting compelling social media narratives around their public actions: infiltrating feminist protest marches, political meetings, the Fashion Week or Cannes Festival; or staging performances wearing burqas or covered with fake blood, in the pure tradition of agit-prop. As well as engineering "buzz" through the repercussions these actions lead to: arrests, fines, altercations... Finally, the group's communicative efficacy is amplified by its articulate and media-genic spokespersons who are all good looking and media-savvy young women.

Despite their repeated assertions that they are not political or ideological actors, Némésis is an integral part of the French and pan European far-right political ecosystem. Founded in 2029, Némésis gained visibility in 2022 during the French presidential campaign, publicly endorsing Éric Zemmour, leader of the far-right party Reconquête, known for its ethnonationalist, anti-immigrant, Islamophobic, and anti-gender agenda. Active figures of Némésis also maintain multiple ideological, personal, and institutional ties with the Rassemblement National. Furthermore, the Collectif Némésis actively engages with a transnational network of women operating on the far-right and identitarian spectrum. The individuals in this network, including influencers, journalists, activists, and politicians, articulate a wide range of grievances, from anti-immigration and anti-Islam rhetoric to anti-trans and ultra-Catholic. Despite this diversity of focus, their efforts coalesce around a common ideological goal: the defense of what they frame as imperiled traditional values concerning Women and the Nation. To consolidate this alliance, Némésis strategically promotes the concept of an "European Sisterhood," aiming to forge a unified front across national borders. Their expansionist strategy takes shape through the establishment of Némésis branches in various countries such as Switzerland, Belgium, Italy, and Spain.

Goddesses of vengeance: weapons and sisterhood

The collective's name itself references the Greek goddess of vengeance, and its logo uses identitarian symbols, notably the Gallic helmet, signaling continuity with French nationalist imagery. In invoking the imagery of its mythological namesake, Némésis portrays the plight of Western women as a crisis demanding retributive action. Framed as a battle against the so-called cultural hegemony of the left and gender theory, Némésis shares the transnational far right's "culture war" (Phelan, 2025; Bleich et al., 2025; May, 2016). This war is performed and enacted through dramatic public actions and through symbolic, at times physical, violence aimed at progressive movements and their supporters as well as immigrant minorities.

Némésis puts forward a powerful image of a fierce warrior ready for battle and their calls to actions are barely veiled. This can be seen in the content shared by members of Némésis on social media. Cordier is for instance seen in 2021 in a video at a shooting range saying 'Bye Bye Sleeping Giants. We're coming' in a video that could not be more explicit, the Sleeping Giants being a citizen collective fighting against the financing of hate speech and one of the main targets of Némésis.



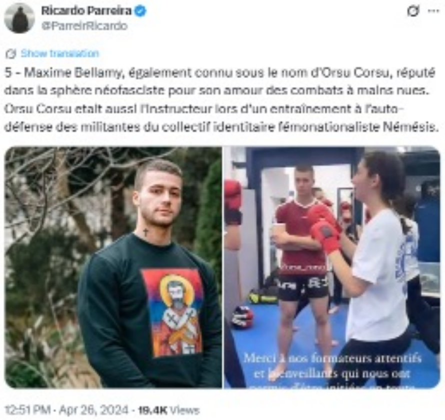
Sleeping Giants FR
@slpng_giants_fr

Translate post

Alice Cordier, porte-parole du collectif xénophobe #Nemesis nous menace de mort par arme à feu. Au nom de la liberté d'expression.
Signalement à la police fait.
Où s'arrêtera l'escalade de la violence d'extrême-droite ?
#LoveNotHate💕 nous continuons !
#Tousmenacés cc @cybergend

5:02 PM · Dec 8, 2021

The activities of Némésis extend beyond imagery to include practical training, as evidenced by a self-defense course they organized in 2023. The instructor for this course was none less than Maxime Bellamy, a French MMA fighter deeply embedded in the violent far-right milieu, where he is recognized as a key member of the neo-Nazi group Les Zouaves and a known hooligan.



Ricardo Parreira
@ParreirRicardo

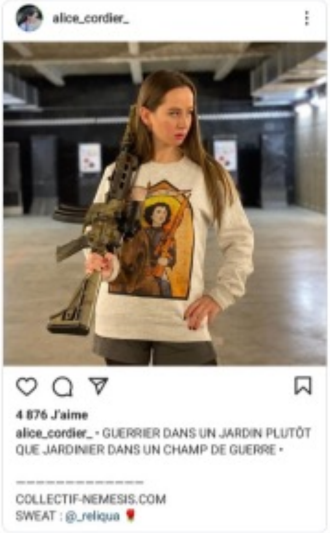
Show translation

5 - Maxime Bellamy, également connu sous le nom d'Orsu Corsu, réputé dans la sphère néofasciste pour son amour des combats à mains nues.
Orsu Corsu etait aussi l'instructeur lors d'un entraînement à l'auto-défense des militantes du collectif identitaire fémonationaliste Némésis.

Merci à nos formateurs attentifs
et bienveillants qui nous ont

12:51 PM · Apr 26, 2024 · 19.4K Views

In another instance, we can see Alice Cordier posing with a machine gun with the quote: « Better be a warrior in a garden than a gardener on the battle field », an adage meaning that it is better to be prepared to fight in time of peace than being unprepared in time of war.



alice_cordier_

4 876 J'aime

alice_cordier_ • GUERRIER DANS UN JARDIN PLUTÔT QUE JARDINIER DANS UN CHAMP DE GUERRE •

—————————————

COLLECTIF-NEMESIS.COM

SWEAT : @_reliqua 🌹

In this picture, she is seen wearing a sweatshirt bearing the image of an armed Joan of Arc, whom she considers one of her role models in terms of feminism. Cordier is here endorsing a product from a label called Reliqua. Popular within the identitarian, far-right and neo-nazi activist milieu (including Maxime Bellamy), Reliqua's products feature imagery of Christian and historical figures bearing weapons (Virgin Mary, Saint Louis…). The design and aesthetic bear a strong resemblance to the Kalashnikov Madonna or Saint Javelin created by Chris Shaw in 2012 in the post-9/11 context. The Collectif Némésis itself has extended its ideological project through merchandising, including sweatshirts emblasoned with the slogan "European Sisterhood." The brand's name consciously positions it within a pre-existing identitarian ecosystem. This nomenclature is strategically deliberate, directly mirroring and feminising the name of the pre-existing identitarian clothing brand European Brotherhood (Keilty, 2018), used by neo-Nazis, especially in Germany. This lexical parallel serves to position Némésis within the same pan-European, ethno-nationalist sphere, while simultaneously articulating a distinct gendered identity within it.

The strategy of Némésis, exemplified by its 'European Sisterhood' clothing collection, operates as a form of metapolitical engagement (Teitelbaum, 2020). This practice of embedding politics into apparel is a key tactic in the contemporary far-right's playbook (Miller-Idriss & Graefe-Geusch, 2020; Gaugele, 2019; Miller-Idriss, 2018).

This fearless image of warriors is cultivated during Némésis' public actions that are meticulously documented on social media. These actions are consistently enabled by its embeddedness within a network of militant far-right groups, which at times provide the security and logistical support necessary for its grassroot operations. This collaboration demonstrates a pattern of mutual aid within the French extremist milieu, where roles are specialised but objectives are aligned. A clear example of this dynamic occurred during the November 20, 2021, demonstration against violence against women in Paris. Némésis inserted itself into the march brandishing a deliberately Islamophobic banner. The group was protected by a security cordon composed of members from identitarian factions such as Action Française, Zouaves Paris, and Cocarde Étudiante, who were armed with various weapons and projectiles (Fondation Jean Jaurès, 2022). This same model of collaboration has been documented elsewhere, including at political rallies for Éric Zemmour (Bouchet, 2021; Libération, 2021). The pattern persisted into 2025, when a known neo-Nazi from the group Division Martel was observed acting as security for Némésis during a March 8 demonstration (StreetPress, 2025).

The components of this network each play a distinct role, beginning with Action Française, a far-right monarchist movement that serves as an ideological and tactical incubator for Némésis. Cordier received her foundational activist training within this organization and acknowledges its lasting influence (Marianne, 2022), which is actually not that surprising as women have been historically involved (Cleret, 2013). Although it presents itself today as a cultural and political group, its youth wing remains militant, with members involved in violent street clashes, assaults, and intimidation — particularly against left-wing and anti-fascist groups on university campuses and during demonstrations in recent years. While the organisation officially denies promoting violence, its activist culture continues to generate physical confrontations and a reputation for aggressive far-right activism. Furthermore, the far-right student union La Cocarde Étudiante, active in over twenty-two universities, provides a recruitment pool and logistical base, with its members consistently identified within the security details protecting Némésis during its infiltrations of and more or less legal participations to feminist marches. Finally, the now-dissolved neo-Nazi group Zouaves Paris specialised in providing muscle as a security force for far-right rallies, including for Némésis (Fondation Jean-Jaurès, 2022). Zouaves's collaboration with other factions, such as providing security for the annual "Comité du 9 Mai" march in honor of Joan of Arc further illustrates the interconnected nature of this ecosystem (République Française, 2022; Le Monde, 2022) in which Némésis is clearly embedded.

The Escalating pattern of confrontational activism of the Némésis Collective

The trajectory of Collective Némésis is marked by an escalating pattern of confrontational activism, which has increasingly attracted formal legal scrutiny. From disruptive actions to formal investigations, the group's legal entanglements illustrate how they move from online provocation to becoming a recurrent subject of criminal complaints. This evolution is marked by a shift from being online commenters to being perpetrators of public disturbance and respondents in investigations and charges for hate speech, defamation, and incitement to hatred.

The year 2024 proved to be a significant period of escalated legal pressure. The group's strategy of disrupting local public events triggered a series of official responses. Following their disruption of a carnival in Besançon, the city's mayor, Anne Vignot, filed a legal complaint for "incitement to racial hatred" (L'Est Républicain, 2024) and "defamation," the latter charge stemming from rhetoric that associated her with "rapists," which she denounced as "extremely violent" (L'Est Républicain, 2025). Vignot then filed another complaint, reporting cyberbullying against her, threats of rape and incitement to commit crimes against her made by Némésis (Le Parisien, 2024). This politician is far from being the only victim: Némésis systematically provokes or is complicit in the harassment of left-wing politicians and feminist activists. The group has developed methods to incite coordinated harassment against its targets while maintaining a level of deniability, making it difficult to be held directly responsible for the ensuing waves of hate (Moissac au Cœur, 2025).

Multiple studies along the years (Walther, 2022; Stahel & Baier, 2023; Näsi et al., 2015) have built a compelling case that experiencing online hate—directly or indirectly—has a clear negative impact on an individual's mental health and overall well-being. More recent studies indicate also that individuals targeted by online hate speech often experience heightened "feelings of insecurity" that impact their life in offline environments (Dreißigacker et al., 2024). This is visible in the case of this elderly activist who in May 2024 interrupted a public poster-pasting action by Némésis collective filed two formal complaints: one specifically against Némésis and another against unknown persons (plainte contre X). The individual was filmed and subsequently led to a severe campaign of harassment, death threats, and a bounty placed on his head. As a direct result, he was compelled to take extreme protective measures: he closed his social media accounts, shaved his moustache, temporarily changed his residence, describing his life as being lived in a state of "permanent stress" (Blaise, 2024).

While tactics like coordinated harassment, swarming, and doxxing are unambiguous crimes, groups like Némésis strategically avoid legal repercussions by operating in a gray zone of hate incitement. Their power lies not in direct commands but in a rhetoric built on plausible deniability (Marks & Stanfill, 2025; Ma, 2024; Hodge, 2020). This involves using strategic ambiguity and symbolic violence—such as labeling opponents as legitimate targets—without issuing explicit orders. They thus mobilise their followers through vilification, all while maintaining a veneer of deniability regarding the resulting real-world harm, such as swarming and threats executed by their base. This insulates the leadership from accountability, even as their discourse directly fuels the actions of their followers.

Discussion

The case of Némésis illustrates that online hate speech is far more than mere digital rhetoric; it functions as a core mechanism of contemporary exclusionary politics. Their online activity, which translates directly into coordinated street actions, helps construct a shared sense of community that exceeds the boundaries of the group itself. In doing so, Némésis embeds its activism within a broader identitarian far-right ecosystem that extends not only beyond national borders but contributes to the formation of a pan-European identity.

Hate speech towards various targets - left feminists, immigrants, media… - is at the core of their discourses and actions. However, one should note, that like many who engage in political activism, far-right activists are also subjected to hate speech and violence—a point that Némésis continually underscores. Indeed, its members are also targeted with online hate speech by far-left activists and politicians, which at times translates into physical violence in the real world. These attacks directed at Némésis members are fully deserving of condemnation.

Notwithstanding this reality, Némésis still clearly operates a renversement de valeurs: while being purveyors - and inciters - of hate speech, they expertly reposition themselves as its primary victims. By claiming that they are the true targets of a pervasive hate and violence from the far-left, they perform a moral inversion. This allows them to deflect criticism, frame any opposition as persecution, and seize the moral high ground traditionally occupied by marginalised groups.

By asserting that they stand on the side of "Good" - a term they actually use - Némésis implicitly constructs its opponents as "Evil," thereby establishing a rigid moral binary. Such Manichean framing recasts conflict in absolutist terms and legitimates the use of sometimes radical means deemed necessary to ensure the triumph of the purportedly "Good" camp. This positioning has been significantly reinforced following the murder of the far-right influencer Charlie Kirk. Némésis posted extensively on the event and participated in a Paris rally organized by a coalition of right-wing and far-right actors, even taking the stage to deliver speeches. In doing so, Némésis and allied groups strategically leveraged the assassination to also portray themselves as martyrs for free speech. They underscored the death threats they themselves receive, implicitly suggesting that a similar tragedy could occur in France. By co-opting the "Je Suis Charlie" slogan and casting themselves as brave combatants in a culture war, they transformed an external tragedy into a potent internal rallying cry. Ultimately, this posture of victimhood and bravery serves to justify potential violence not as aggression, but as a necessary and legitimate form of self-defense. The discourse of free speech thus becomes a shield, insulating them from accountability and reframing retaliatory actions as the brave duty of those under siege.

References

Akkerman, T. (2015). Gender and the radical right in Western Europe: A comparative analysis of policy agendas. *Patterns of Prejudice*, 49(1–2), 37–60.

Alestra, L. (2025). *Les vigilantes. Surveillées et surveillantes, ces femmes au cœur de l'extrême droite*. JC Lattès.

Almony, L. (2022). Hegemonic masculinity and the contemporary rise of female right-wing populist leaders: The case of Giorgia Meloni. *Journal of Political Inquiry,* 34, 34-47.

Aouidad, K. (2022, September 13). Maxime Bellamy, freefighter, néonazi et hooligan à Rennes.*StreetPress*. https://www.streetpress.com/sujet/1662996573-maxime-bellamy-freefighter-neonazi-hooligan-rennes-militaire-francais-combat-violences-zouaves-extreme-droite-mma

Bacchetta, P., & Power, M. (Eds.). (2002). *Right-wing women: From conservatives to extremists around the world*. Routledge.

Bartlett, J., Krasodomski-Jones, A., & Littler, M. (2019). *Hate in the machine: Anti-Muslim hate online and offline*. Demos.

Berry, V. (2012). Ethnographie sur Internet : rendre compte du « virtuel ». *Les Sciences de l'éducation. Pour l'Ère nouvelle*, 45(4), 35-58. https://doi.org/10.3917/lsdle.454.0035.

Blaise, L. (2024, November 19). "On va te mettre deux balles dans la tête" : depuis qu'il a interrompu une action d'extrême droite, il vit un "stress permanent". *France 3 Bourgogne-Franche-Comté.* https://france3-regions.franceinfo.fr/bourgogne-franche-comte/doubs/besancon/on-va-te-mettre-deux-balles-dans-la-tete-depuis-qu-il-a-interrompu-une-action-d-extreme-droite-il-vit-un-stress-permanent-2977064.html

Bleich, E., Zsombok, G., & van der Veen, A. M. (2025). Social proximity, discursive opportunity structures, and the diffusion of the culture wars: The case of "woke" in France. *International Journal of Comparative Sociology*, 00207152251358849.

Bonnefond, C. (2024, July 3). *Les femmes votent-elles plus à gauche que les hommes ? Pas dans la France de 2024*. CNRS. https://www.inshs.cnrs.fr/fr/cnrsinfo/les-femmes-votent-elles-plus-gauche-que-les-hommes-pas-dans-la-france-de-2024

Bouchet, G. (2021, December 6). Néonazis, royalistes, identitaires : bienvenue au meeting de Zemmour. *Politis*. https://www.politis.fr/articles/2021/12/neonazis-royalistes-identitaires-bienvenue-au-meeting-de-zemmour-43865/

Calderaro, C. (2025). Beyond Instrumentalization: Far-Right Women's Appropriation of Feminism in France. *Politics & Gender*, 1-29. https://doi.org/10.1017/S1743923X24000412

Camacho, A. N. (2021). Right-wing feminism and the securitization of migration: On the example of the german campaign 120 Dezibel. *InterNaciones*, (21), 91-110.

Casez, J. (2024, May 30). "On va te mettre deux balles dans la tête" : depuis qu'il a interrompu une action d'extrême-droite, il vit "un stress permanent". *France 3 Bourgogne-Franche-Comté*. https://france3-regions.francetvinfo.fr/bourgogne-franche-comte/doubs/besancon/on-va-te-mettre-deux-balles-dans-la-tete-depuis-qu-il-a-interrompu-une-action-d-extreme-droite-il-vit-un-stress-permanent-2977064.html

Cleret, C. (2013). De la charité à la politique: l'engagement féminin d'Action française. Parlement [s], *Revue d'histoire politique*, 19(1), 17-29.

CNews. (2025, January 23). La directrice du collectif Némésis, Alice Cordier, annonce porter plainte contre la « nébuleuse StreetPress ». *CNews* https://www.cnews.fr/france/2025-01-23/la-directrice-du-collectif-nemesis-alice-cordier-annonce-porter-plainte-contre-la

Critchlow, D. T. (2018). Phyllis Schlafly and grassroots conservatism: A woman's crusade. In *Phyllis Schlafly and Grassroots Conservatism*. Princeton University Press.

Debras, F. (2022, August). *The ambivalence of a far-right feminism: political strategy or ideological reconfiguration of the gender question? A Critical Discourse Analysis of the Rassemblement National.* In ECPR General Conference, University of Innsbruck, 22.

Della Sudda, M. (2025). Succès et infortune de la rhétorique réactionnaire des alterféministes écolo-conservatrices. *Bifurcation/s: la revue des écologies politiques émancipatrices*, (2).

Della Sudda, M. (2022). *Les nouvelles femmes de droite*. Hors d'atteinte.

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024). Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science,* 13(1), 4.

European Union. (2022). Regulation (EU) 2022/2065 on a single market for digital services (Digital Services Act). *Official Journal of the European Union*, L 277, 1–102.

Eksi, S. (2021, October 29). *Far right female extremism and leadership: Their power of framing reality in the European context*. European Centre for Populism Studies. https://www.populismstudies.org/far-right-female-extremism-and-leadership-their-power-of-framing-reality-in-the-european-context/

Farris, S. R. (2017). *In the name of women's rights: The rise of femonationalism.* Duke University Press.

Fondation Jean-Jaurès. (2022). *Les poings sans la rose : Les groupuscules d'extrême-droite en campagne en 2022*. https://www.jean-jaures.org/publication/les-poings-sans-la-rose-les-groupuscules-dextreme-droite-en-campagne-en-2022/

France Bleu. (2025, February 13). À Caen, une étudiante transgenre porte plainte contre le collectif Némésis. *FranceBleu*. https://www.francebleu.fr/infos/faits-divers-justice/a-caen-une-etudiante-transgenre-porte-plainte-contre-le-collectif-nemesis-2243707

François, S. (2021). *La Nouvelle Droite et ses dissidences. Identité, écologie et paganisme.* Bord de l'eau.

Gaugele, E. (2019). The new obscurity in style. Alt-right faction, populist normalization, and the cultural war on fashion from the far right. *Fashion Theory,* 23(6), 711-731.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation,* *and the Hidden Decisions That Shape Social Media*. Yale University Press.

Goetz, J. (2022). 'Patriotism is not just a Man's Thing': Right-wing Extremist Gender Policies within the so-called Identitarian Movement. *Journal of modern European history*, 20(3), 389-406.

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media black box: Digital platforms and content moderation. *Big Data & Society,* 7(1). https://doi.org/10.1177/2053951720928996

Hemmer, N. (2016). *Messengers of the right: Conservative media and the transformation of American politics*. University of Pennsylvania Press.

Hine, C. (2015). *Ethnography for the Internet: Embedded, Embodied and Everyday*. Bloomsbury.

Hodge, A, (2020). Plausible deniability in McIntosh, J., & Mendoza-Denton, N. (Eds.). *Language in the Trump era: Scandals and emergencies*. Cambridge University Press.

Jouët, J., & Le Caroff, C. (2013). Chapitre 7-L'observation ethnographique en ligne. *Collection U*, 147-165.

Kay, J. B. (2024). The reactionary turn in popular feminism. *Feminist Media Studies,* 1-18.

Keilty, P. (2018). Pornography's White Infrastructure. *Catalyst: Feminism, Theory, Technoscience* 4(1): 1-9

Köttig, M., Bitzan, R., & Petö, A. (Eds.). (2017). *Gender and far-right politics in Europe.* Palgrave Macmillan.

Kress, G., & van Leeuwen, T. (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Arnold.

Le JDD. (2023, June 15). Le collectif féministe identitaire Némésis défie les menaces et s'étend en Europe. *Le Journal du Dimanche*.

Le Monde. (2024, February 8). *Némésis, collectif xénophobe et « féministe » autoproclamé*. https://www.lemonde.fr/campus/article/2024/07/05/nemesis-collectif-xenophobe-et-feministe-autoproclame-et-ses-batailles-ideologiques_6247219_4401467.html

Le Monde. (2022, January 5). *Le groupuscule d'ultradroite Les Zouaves Paris dissous en conseil des ministres*. https://www.lemonde.fr/societe/article/2022/01/05/le-groupuscule-d-ultradroite-les-zouaves-paris-dissous-en-conseil-des-ministres_6108294_3224.html

Le Parisien. (2024, April 13). *Des centaines d'injures: la maire de Besançon harcelée en ligne après sa plainte contre les pancartes anti-migrants*. https://www.leparisien.fr/faits-divers/des-centaines-dinjures-la-maire-de-besancon-harcelee-en-ligne-apres-sa-plainte-contre-les-pancartes-anti-migrants-13-04-2024-C6HZMFFCIJGQXP6UJ2PL22GF7M.php

L'Est Républicain. (2024a, April 7). *Némésis perturbe le carnaval, Anne Vignot va déposer plainte pour incitation à la haine raciale*. https://www.estrepublicain.fr/faits-divers-justice/2024/04/07/nemesis-perturbe-le-carnaval-anne-vignot-va-deposer-plainte-pour-incitation-a-la-haine-raciale

L'Est Républicain. (2024b, May 20). *Après Besançon, le collectif identitaire Némésis perturbe le festival Cirque et Fanfares à Dole*. https://www.estrepublicain.fr/faits-divers-justice/2024/05/20/apres-besancon-le-collectif-identitaire-nemesis-perturbe-le-festival-cirque-et-fanfares-a-dole

L'Est Républicain. (2025, January 9). *Le collectif d'extrême droite Némésis perturbe la cérémonie des vœux aux agents, Anne Vignot dépose plainte*. https://www.estrepublicain.fr/faits-divers-justice/2025/01/09/le-collectif-d-extreme-droite-nemesis-perturbe-la-ceremonie-des-voeux-aux-agents-anne-vignot-depose-plainte

Libération. (2021, November 21). *Comment les militantes identitaires de Némésis ont perturbé la manif #NousToutes*. https://www.liberation.fr/politique/comment-les-militantes-identitaires-de-nemesis-ont-perturbe-la-manif-noustoutes-20211121_6QMOJEXYF5GDZDYULHD3FKNRA4

Ma, C. (2024). Overcoming Far-Right Respectability: The Case for Systemic Approaches to Studying White Supremacy. *Political Communication*, 41(6), 1035-1040.

Marianne. (2022, May 2). Qui est Alice Cordier, cette militante d'extrême droite à l'origine des révélations sur Ersilia Soudais? *La Marianne* https://www.marianne.net/societe/qui-est-alice-cordier-cette-militante-dextreme-droite-a-lorigine-des-revelations-sur-ersilia-soudais

Marks, R., & Stanfill, M. (2025). The plausible deniability playbook: how white victimhood narratives evade moderation. *Communication and Critical/Cultural Studies*, 1-18.

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Polity Press.

Masquelier, J. (2015). «Pour un genre catholique!» Trajectoire de l'association Femmes et Hommes dans l'Église (1970-2000). Sextant. *Revue de recherche interdisciplinaire sur le genre et la sexualité*, (31), 43-58.

May, P. (2016). French cultural wars: public discourses on multiculturalism in France (1995–2013). *Journal of Ethnic and Migration Studies*, 42(8), 1334-1352.

Mediapart. (2022, April 27). *Némésis: ces féministes identitaires au service du camp national*. https://www.mediapart.fr/journal/france/270422/nemesis-ces-feministes-identitaires-au-service-du-camp-national

Miller-Idriss, C., & Graefe-Geusch, A. (2020). Iconography and embodiment in far right youth. *Researching the Far Right: Theory, Method and Practice*.

Miller-Idriss, C. (2018). *The extreme gone mainstream: Commercialization and far-right youth culture in Germany*. Princeton University Press.

Moffitt, B. (2016). *The global rise of populism: Performance, political style, and representation*. Stanford University Press.

Mudde, C. (2007). *Populist radical right parties in Europe*. Cambridge University Press.

Mudde, C., & Rovira Kaltwasser, C. (2017). *Populism: A very short introduction*. Oxford University Press.

Moissac au Cœur. (2025, August 20). *Némésis, le collectif d'extrême droite qui provoque le cyberharcèlement de militantes féministes et d'élues de gauche*. https://moissacaucoeur.fr/2025/08/20/nemesis-le-collectif-dextreme-droite-qui-provoque-le-cyberharcelement-de-militantes-feministes-et-delues-de-gauche/

Msika, P. (2025). Les élections législatives de 2024 et le retour à la réalité. *Le Droit de Vivre*, 694(1), 68–70. https://doi.org/10.3917/ddv.694.0068

Mudde, C., & Kaltwasser, C. R. (2017). *Populism: A very short introduction*. Oxford University Press.

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association,* 19(4), 2131–2167. https://doi.org/10.1093/jeea/jvab014

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

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press.

Pernes, C. (2025, 7 March). Collectif d'extrême droite Némésis : "Elles utilisent la cause des femmes pour faire avancer la cause nationaliste, identitaire et sécuritaire". *Télérama*. https://www.telerama.fr/debats-reportages/collectif-d-extreme-droite-nemesis-elles-utilisent-la-cause-des-femmes-pour-faire-avancer-la-cause-nationaliste-identitaire-et-securitaire-7024655.php

Phelan, S. (2025). Seven theses about the so-called culture war (s)(or some fragmentary notes on 'cancel culture'). *Cultural Studies,* 39(1), 63-88.

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & Tacchi, J. (2016). *Digital Ethnography: Principles and Practice*. Sage.

Plottu, P. & Macé, M. (2024). Chapitre 4. Le rôle des femmes. Pop fascisme : Comment l'extrême droite a gagné la bataille culturelle sur internet. (p. 89-

103). *Divergences*. https://shs.cairn.info/pop-fascisme--9791097088743-page-89?lang=fr.

Politico. (2024, June 10). How France's far right won over women voters. *Politico*. https://www.politico.eu/article/france-eu-elections-2024-women-vote-far-right-policy-emmanuel-macron-july-7/

Raison du Cleuziou, Y. (2019). *Une contre-révolution catholique. Aux origines de La Manif pour tous.* Éditions du Seuil.

République Française. (2022, January 5). Décret n° 2022-11 du 5 janvier 2022 portant dissolution d'un groupement de fait. Legifrance. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044844463

Rey, S., & Gonthier, F. (2024, October 21). L*es jeunes femmes plus à gauche que les jeunes hommes ? Le vote aux élections législatives de 2024 au prisme du genre et de l'âge* [Conference presentation]. Institut national d'études démographiques (INED). https://www.ined.fr/fr/actualites/rencontres-scientifiques/les-lundis/les-jeunes-femmes-plus-a-gauche-que-les-jeunes-hommes-le-vote-aux-elections-legislatives-de-2024-au-prisme-du-genre-et-de-l%20age/

Rydgren, J. (2007). The sociology of the radical right. *Annual Review of Sociology*, 33, 241–262.

Rydgren, J. (Ed.). (2018). *The Oxford handbook of the radical right*. Oxford University Press.

Scheyer, V., True, J., & Bonotti, M. (2025). Analysing power and gender dynamics in the German far-right network: a feminist approach to social network analysis. *European Journal of Politics and Gender*, 1(aop), 1-28.

Scrinzi, F. (2017). "Gender and the European Far Right." In T*he Oxford Handbook of the Radical Right*, edited by J. Rydgren, Oxford University Press.

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

StreetPress. (2025, March 8). *À la manif du 8 mars, Némésis défile avec un néonazi du groupe Division Martel.* https://www.streetpress.com/sujet/1741788081-manifestation-nemesis-8-mars-neonazi-service-ordre-division-martel-sarah-knafo-agression-feminisme-femonationalisme

Taggart, P. (2000). *Populism*. Open University Press.

Teitelbaum, B. R. (2020). *War for eternity: The return of traditionalism and the rise of the populist right*. Penguin Books.

Teitelbaum, B. R. (2017). *Lions of the north: Sounds of the new Nordic radical* *nationalism*. Oxford University Press.

Tichenor, K. A. (2014). *Protecting Unborn Life in the Secular Age: The Catholic Church and the West German Abortion Debate, 1969–1989*. Central European History, 47(3), 612-645.

Uzelac, M. (2023). Algorithmic governance of hate speech in the EU. *European Journal of Communication*, 38(2), 149–166. https://doi.org/10.1177/02673231221143239

Vendrand-Maillet, B. (2020). La gauche face à la nouvelle génération. *Le Débat,* 209, no. 2, 109-119.

Venner, F. (1993). Le militantisme féminin d'extrême droite: «Une autre manière d'être féministe»? *French Politics and Society*, 11(2), 33–54.

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology,* 45, 101298. https://doi.org/10.1016/j.copsyc.2021.12.010

Williams, M. L., & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2), 211–238. https://doi.org/10.1093/bjc/azv059

Social media:

Đào, L. [@linhlandao]. (2022, November 25). Le collectif Némésis, qui se présente comme un mouvement de femmes, est financé par des hommes milliardaires d'extrême droite [Image attached] [Post]. X. https://x.com/linhlandao/status/1596140486280237056

Sleeping Giants FR [@slpng_giants_fr]. (2021, December 8). Le collectif Némésis se présente comme un "mouvement de femmes" qui veut "réenchanter la France". Mais qui [Image attached] [Post]. X. https://x.com/slpng_giants_fr/status/1468606141949222917

Let’s not Talk about Sex – Anti-Gender Coalitions and Fears of “Demonic” Sex Education in Belgium

Katrien Jacobs & Katelijne Lievens

Introduction: Coherent Storylines and Gender Phantasms

In order to understand how different political, cultural, and religious groups became “unlikely bedpartners” and united around their fear of “gender,” it is useful to think of their immersion in “discourse coalitions,” or how they sustained a massive social hysteria based on an emotively coherent story line. Hajer has defined discourse-coalitions as “groups of actors that, in the context of an identifiable set of practices, share the usage of a particular set of story lines, over a particular period of time” (Hajer 2006, 70). Not only can actors conceal the discursive complexity of the subject, but the use of story lines also creates an illusion of oneness - even though the actors might have disparate ideological backgrounds (Hajer 2006; Hajer 1993, in Evans et al. 2021, 2). In these story lines, actors often make use of “emotive metaphors,” symbolic terms or images that make abstract concepts easily understandable and powerful. The anti-gender coalitions focused on the “vulnerability” of those people who needed to be protected the most: the children. By putting the emphasis on the sexual harassment of children, they turned the protection and the education of the children into a symbol for the broader societal tension. In addition, the protection of children was framed as a problem that needed to be solved by a clear moral duty. Repeated slogans such as “unite to protect children,” and “hands off our children” were present amongst many groups, creating an echo-chamber like effect (Cinelli et al. 2021, 1; Gillani et al. 2018, 823; Terren and Borge 2021, 100-101). The creation of such “discourse coalition” around the sexual harassment of children was facilitated by digital platforms and algorithms, which enabled activists to foster collaboration and coordinate action beyond the immediate social circles (Evans et al. 2021, 6; 19). In our case, we mostly analyzed anti-gender discourses on X, the platform that was purchased by Elon Musk in 2022 and whose moderators and algorithms are already known to support actors in favor of white nationalism and gender conservatism.

In *Who’s Afraid of Gender* (2024), Butler argues that unlikely coalitions of disparate actors are built around shared “gender phantasms,” imaginary constructs that condense and string together a wide range of moral panics, conspiracies and fears about the society at large. Different psychic and social ills are arbitrarily connected, and "gender" is reduced to a single, monolithic and monstrous entity. They are also connected to a "decay-phantasm," which posits gender as an elitist, idly academic and destructive force able to pulverize the pillars of Western society and the heteronormative family. As Butler's and Syed's lecture (2024b) about Anti-Gender Movements shows, they thrive on an excited sense of moralism and a desire to fix the decaying world by means of cruelty, disdain, resentment and destruction. The decay-phantasm dictates that gender has become destructive and demonic and hence needs to be eliminated alongside the field of gender studies, gender studies scholars, health care, kinship, transrights, LGBTQIA+ rights, erotica and nudity. It is not an easy task to dismantle these phantasms or to come up with alternative non-rationalist sources that would be equally able to motivate or inspire people. Not only do the groups make use of coherent story lines and emotive issues, but they also posit easily digestible metaphors to make their ideology more tangible.

In our case-studies, both the fear of encroaching "demonic" sex education and the genderbread cookie became a driving force for this movement. In both case-studies, we observed discourse-coalitions the year preceding Belgium's Federal Elections, which took place in June 2024. We did a qualitative ethnographic analysis by observing actors and by sampling and comparing posts that were shared during the Walloon and Brussels attack on sex education and the Flemish revulsion of "genderkoek". We also observed cross-border collaborations and synergies between The Dutch and Flemish far-right, as the Dutch far-right leader Geert Wilders visited Belgium during the 2024 federal elections and his previous attacks on gender were recycled as one of the main political campaigns. We chose to focus on two separate case-studies taking place in the Walloon and Flemish region of Belgium respectively, each with a different focus on sex/gender politics but both stirred by similar "gender phantasms." The first incident was unprecedent violent attack in French-speaking Belgium on progressive sex education that was clearly organized by grassroots agents. The second incident was part of a larger far-right election campaign in Dutch-speaking Belgium that had been pioneered by far-right leaders in the Netherlands. Paternotte and Kuhar state that it is important to look at localized anti-gender movements to point out their cultural differences (Paternotte and Kuhar 2018, 7). At the same time, our essay shows that anti-gender movements in Belgium are turning into heterogeneous and amorphous coalitions that are steered by online misinformation.

EU Anti-Gender Movements and Coalition

Anti-gender manifestations are ubiquitous in contemporary Europe and several informal coalitions have emerged that take activist positions against "gender ideology" (Graff and Korolczuk 2021). In this specific context, the opponents see gender as a compilation of ideas that go against both science and religion (Butler 2024, 4-5). This opposing coalition consists of many different groups of diverse cultural backgrounds, such as the white nationalist far-right, conservative Christian and Muslim religious groups, and conspiracy followers. Generally speaking, gender ideology is considered to be an evil monolith and a hidden political conspiracy, which tries to seize power and force certain values onto "ordinary people".

Anti-gender movements promote highly emotive claims and misinformation about gender, but they do have the intention to have actual impact in society by influencing legislation such as intimate/sexual citizenship, LGBTQI+ rights, reproductive rights, as well as sexual and gender education (Butler 2024, 4-5; Paternotte and Kuhar 2018, 8-12). Especially through the use of shared discourses on social media and online platforms, these coalitions can organize themselves without having to meet in real life, creating a shared philosophy across many different backgrounds (Norris 2023, 17-18). As a result of these anti-gender coalitions, the EU's pioneering policies over LGBTQI+ right and sex education are now in grave danger of being pushed back (Datta 2018, 27). The Christian roots of the anti-gender movement can be traced back to the mid-90s, when the Catholic church resisted the UN Conferences of 1994 and 1995. A few years later, in the mid 2000s, conservative groups and political parties fought against the law regulating gay marriage (Paternotte and Kuhar 2018, 7). By changing the meaning of "gender" from a neutral term to an ideological matrix, the Vatican was able to unite different actors in opposition against feminism, gender and liberalism. Despite the fact that not everyone who uses anti-gender rhetoric identifies as Catholic, the Vatican has thus been of pivotal importance in creating a framework for this ideology to arise (Paternotte and Kuhar 2018, 11-12).

The Catholic Church also emphasized that gender ideology and the acceptance of "endless" gender identities was a kind of "sexual perversion." For example, the work of Sister Marguerite Peeters of the Brussels NGO Intercultural Dialogue Dynamics proposed a similar anachronistic body. She is an advisor to Pope Francis' pontificate and wrote that gender studies are typical of the "mystery of evil". The contestation of gender as binary would be nothing less than a "threefold perversion," a disorderly quest for power, for pleasure and for knowledge, types of life pursuits that forefront pleasure in and of itself. She called it a project of destructive alternative knowledge that plots against "normal" humanity and attempts to impose minority values. Peeters raged against the ad infinitum inventions and chaotic phantasms of identity politics:

> Indeed, the gender theory and its natural extension – the queer theory which goes as far as affirming that the male or female body is a social construct – put a great strain on reason. The gender theoreticians fight among themselves over the meaning of the expressions they themselves forged, such as sexual identity, gender identity, sexual norms, sexual orientation or preference, sexual role, gender role, sexual behavior, gender stereotype, sexual diversity and so on. The proliferation of lexicons attempting to clarify ad infinitum the specificities of the numerous expressions declining the gender concept only strengthens the Babel tower in which we live, as they often contradict themselves. (Pontifical Council of the Laity, 2011)

In addition to the Catholic Church, far-right populism played an important role in several anti-gender movements. In certain countries, such as Poland and Italy, far-right populists were the main drivers behind the anti-gender protests. Throughout Europe, the organization *Agenda Europe* became a pioneer of anti-gender activism. In 2017, they published their manifesto, *Restoring the Natural Order: an Agenda for Europe* (Datta 2018, 5). According to this organization, there would be a natural order that cannot be changed. This order would stipulate that gender is fixed, and that the only purpose of sexual activities should be procreation (Datta 2018, 17). *Agenda Europe* was pivotal in mainstreaming anti-gender discourses and emotions in Europe. They also relied upon the classic blueprint of fascism. Particularly the "never ending war", "a mythical past", and the "inherently violent nature of the people" were translated into their manifesto in order to limit the "unnatural freedom" of people and to return to a past that in fact never existed (Norris 2023, 16; 74-75). Agenda Europa also argued that the natural order of society was threatened by "demonic" revolutions, which are further associated with sexual and erotic pleasures, abortion lobbies, LGBTQIA+ activism, feminisms and militant atheism (Datta 2018).

Case study 1: Attacks on EVRAS Sex Education (Fall 2023)

L'Éducation à la Vie Relationnelle, Affective & Sexuelle (also known as EVRAS) was created on the 12th of July 2012 in Belgian schools of the French-speaking community and a protocol was signed to ensure its implementation (EVRAS, n.d.a; EVRAS, n.d.b). EVRAS tries to ensure the facilitation of reliable and understandable information to children, and to develop a positive lens regarding sexuality. At the same time, they want to create a critical attitude around the hypersexuality of the environment. EVRAS also constitutes a recognition of the importance of sexual education, in accordance with the recommendations of UNESCO (Unesco Platform Vlaanderen 2018; EVRAS, n.d.b). Based on this protocol, school directors are asked to take initiatives regarding EVRAS and are recommended to work together with specialized centra, such as psycho-medical-social centra or health promotion services. Starting from the school year of 2023-2024, it became mandatory for schools to organize at least one EVRAS-activity of two hours in the sixth year of primary school, and one activity of two hours in the fourth year of secondary school (De Lobel 2023; EVRAS, n.d.b).

As often happens with anti-gender protests, changes in the curriculum of school children's education were a trigger for resistance (Paternotte and Kuhar 2018, 8). It started with an open letter by children's psychiatrist Sophie Dechêne, which was signed over more than 7 500 times (Baumers and Clemens 2023; Vancaeneghem 2023). From there on, a snowball effect was set in motion. An interview with Dechêne was cut up and edited, resulting in conspiracy theories on social media (Baumers and Clemens 2023; Titeca 2023). Despite the different backgrounds of signatories, the perceived problem with EVRAS was that it "teaches children how to masturbate, teaches toddlers that gender surgeries are fun, and it approves of porn" (Debruyne 2023). The ensuing street protests were peaceful, bringing several hundreds of parents from differing political groups together in Brussels in September-October 2023 (BRUZZ 2023a; BRUZZ 2023b; Decré 2023; VRT NWS 2023). But the misinformation about EVRAS also became triumphant in these protests as well. The initial peaceful events spiraled into less-peaceful protest actions, and eventually led to arsons in several schools (Stroobants 2023).

On the 2nd of October 2023, several Islamic organizations announced to go to the constitutional court, in an attempt to further stop EVRAS (Mouhamou 2023; Redactie De Morgen 2023). In the news media, these protests were framed as being led by "veiled mothers" but there were two main driving forces behind the protests, Radya Oulebsir and Nicolas Lefèvre, as well as many smaller actors. Oulebsir is a Muslim mother, mostly known for her YouTube channel with more than 14.000 followers. Additional to this loyal support group, seven Belgian Muslim organizations backed her up (Struys 2023). In her videos, Oulebsir talks about how the government discriminates against parents, who are deemed incapable of talking about sexual education. Not only is there, according to Oulebsir, no disclosure about what is discussed during the classes. Moreover, the EVRAS program would be part of a bigger conspiracy, in an attempt to thin the world population (Baumers and Clemens 2023). On YouTube, she appealed to parents to take their children out of school, so that their children will not be exposed to the nonsense of EVRAS (@ Omrri Omrri, 2023).

The second leader was Nicolas Lefèvre, chairman of *Bon Sens Belgique*, an organization founded in 2020 in response to the Covid-19 crisis. Allegedly, the group is incensed by the way society "lost their common sense." Their goal is to inform and unite people, and to create more clarity about an incomprehensible world (Bons Sens Belgique n.d.). They do this in several ways, such as writing books, publishing the newspaper *Kairos*, and through street protests (Rédaction RTBF 2023). They describe themselves as the defendants of freedom and critical thinking, in an attempt to break with the dominating ideologies in our society (Kairos n.d.). These two main forces were joined by a few other, smaller organizations, amongst others *Zone Libre,* invited by Lefèvre and founded by Daniel de Wolff, an organization with the goal to "protect the Belgian federal law and constitutional state" (La zone Libre n.d.). Not only did they again want to resist EVRAS, but they also rallied against covid-vaccinations as emblematic of the power of "international authorities" (Rédaction RTBF 2023). Civitas Belgique was also present, a French-Belgian nationalist-catholic movement (Struys 2023; De Coninck 2023; Justaert 2023). Their chairman, Alain Escada, also engages in nation-state politics by supporting the far-right Front National Belgique (FNB) (Struys 2023; De Coninck 2023; Justaert 2023).

Lastly, *Sauvons Nos Enfants* was protesting against EVRAS, with as its main motivation the protection of the innocence of children (Sauvons Nos Enfants, n.d.a; Struys 2023). This group was founded by Frederic Goareguer, a pediatric psychiatrist, who has previously been known for spreading fake news during the Covid-19 pandemic (Struys 2023). This collective specifically stated that they would be open to sexual education at school, if there would be unbiased proof that this might benefit the children themselves, and the broader society. They did attack EVRAS, as it was in their view not compliant with the Belgian constitutional law (Sauvons Nos Enfants n.d.a; Sauvons Nos Enfants n.d.b). According to Goareguer, they did not have a problem with LGBTIA+, but with presenting ideology and propaganda to children (Van Maele and De Lobel 2023). The name Sauvons Nos Enfants might also ring a bell; as is the literal translation of one of the most famous slogans of QAnon, the American far-right conspiracy group (Struys 2023; North 2020).

The protests gained attention throughout Belgium, and it did not take long for support to arrive from Flanders. Soon, there also arose campaigns again Sensoa, a Flemish expertise center for sexual health, funded by the Flemish government. Comité Bezorgde Ouders (translated as Committee of worried parents) criticized the Sensoa curriculum, stating that gender ideology would be forced upon young children (Justaert 2023). Zone Libre also started a campaign against Sensoa, to stop "sexual education in kindergarten" (La Zone Libre n.d.). Additionally, far-right politicians were getting involved, as Tom Van Grieken, chairman of the far-right political party Vlaams Belang stated in a talk show:

> "When sexual education teaches children about gender ideology, I would not be satisfied with that as a parent either. I do not believe in gender. […] I would also oppose that [the classes about sexual education]. We [Vlaams Belang] are the leaders of the criticism of all this transgender propaganda" (VRT NWS et al. 2023).

When asked about EVRAS, he made it noticeably clear that Vlaams Belang did not believe in "gender" and their elections campaigns in 2024 would be largely based on anti-gender attacks and disinformation (Santens 2024; Van Bakel, Debackere, and Dorjbayar 2023).

At first, it might seem that these two leaders and the smaller groups, including conservative Muslims, Christians, and the far-right, would be hostile to each other, or at least incompatible. However, they reinforced a coherent story line about sex and gender and became allies of a strong sex-phobic vision, namely that the government should not get involved in the upbringing of their children (Struys 2023). Additionally, it is clear that many groups had already bonded against Covid 19-vaccinations during the pandemic and adhered to conspiracy theories. They were triggered by an all-powerful government, which was forcing rules on the common people, and innocent children.

The conspiracy theories were the icing on the cake, as discourse coalitions emerged that allowed strange bedpartners to successfully organize protests. Hajer's notion of discourse coalition shows us that different actors can work together to contest certain forms of politics, without necessarily having common values (Hajer 2006, in Edenborg 2023, 177). The separate groups are able to create the appearance of unity, as if they form one unified front against the sexualization of children and sex education. (Edenborg 2023, 176). The groups might have very differing backgrounds and views, but they are held together by the creation of a "common understanding of the nature of the problem" (Hajer 1993, in Bossner and Nagel 2020, 312). It would be a moral duty for all these different actors to protect "our" children. This became abundantly clear, as the groups called themselves a "front of consciousness" (De Coninck 2023). The groups were willing to put their initial differences aside in order to maintain common sense and stop the "gender nonsense."

Online, this becomes clear in hashtags such as #Agenda2030 and #TheGreatReset (@DanieldeWolff 2023; @LaZoneLibre 2023). The groups went as far as to say that EVRAS would teach five-year-olds to masturbate and porn would be shown in class to nine-year-olds. This is for example what their ally French rapper ROHFF tweeted, leading to a petition that was signed more than 13.000 times (@rohff 2023). These tweets and online videos relied on highly emotive language, visuals, and misinformation to support their case. They focused on "waking up" and "protecting the children," often in capital letters and with several explanation points to emphasize the urgency of the case. This resulted in tweets such as "réveillez vous!!!" and videos titled "EVRAS: RETIRONS TOUS NOS ENFANTS DE L'ECOLE" (@Omrri Omrri 2023; @rohff 2023). And indeed, schools had already been set on fire so that children could safely stay at home and would not have to attend sexual education.

Case-study 2: Revulsion about The Genderbread Cookie (November 2023 to June 2024) (2000)

International Anti-gender organizations, such as Agenda Europa, the Flemish NGO Comité Bezorgde Ouders (Concerned Parents), as well as far-right leaders also started venting their concerns about LGBTQIA+ activism and focused their anger on the symbol of the "genderbread cookie" (genderkoek in Dutch) and its ability to demonize children and youth. Federal elections took place in Belgium in 2024, in which and the Flemish Far-right party Vlaams Belang issued a campaign that included full-frontal attacks on gender. They adopted the rhetoric of anti-gender movements that dictated that "gender does not exist," and that elastic gender identities, LGBTQIA+ rights and sex/gender education are "typically leftwing" fabrications. Just like in Wallonia and Brussels, the attacks on gender attracted an unlikely coalition of ethnic, political, and religious groups, as well as fervent and outspoken conspiracy theorists. All united around "gender phantasms," or a fear that progressive sex/gender policies would have a negative impact on youth by "hypersexualizing" them (Norris 2023; Butler 2024).

Again, some of these uniting story lines had emerged during the Covid-19 pandemic, as communities of anti-vaxxers and QAnon conspiracy theorists had emerged and showed an obsession with protecting their bodies from toxic interference. In this regard, the Flemish far-right intersected with anti-vaxxers in the Netherlands, many of whom believed that pedophiles are actively operating in the Low Countries and are supported by global political elites. Attacks on the gender break cookie (genderkoek in Dutch) had been initiated in 2021 by Geert Wilders of the Dutch far-right party PVV (Party Voor Vrijheid, or Party For Freedom) and were taken up by Flemish politicians in the weeks preceding Federal elections. They rallied around a perceived indoctrination of youth with LGBTIQIA+ icons, such as the rainbow flag or the genderbread cookie. The genderbread cookie specifically triggered a lot of online outrage, eliciting many calls to defund Belgian organizations who would bring such pedagogical tools to school children and "ram them down their throats." The genderbread cookie was an example of global left-wing elites taking over society and leading a revolution that would destroy society. The Flemish far-right set up small anti-gender NGO Bezorgde Ouders (Concerned Parents) whose primary goal became to attack organizations devoted to sexual rights and health. Agnes Jonckheere, a Christian conservative representative of the far-right Samen voor Democratie (Together for Democracy), a Flemish off-shoot of the Dutch far-right Forum Voor Democratie, (FvD), devoted an entire internet seminar to attacking sex education and the introduction of "gender ideology" in Belgian primary and secondary education (Bezorgde Ouders, 2023). These attacks on progressive sex education then condensed into a coherent sentiment against a "woke" educational tool, the genderbread cookie.

The genderbread cookie was created in 2011 by Sam Killerman as a tool to explain different aspects of gender, such as physical-biological embodiment, identity, desire, and expression. The cookie explains that sexuality and gender can be seen as different layers of human consciousness and that those aspects can be at odds with each other. During the 2024 elections in Belgium, there was a reactionary outburst against this cookie, as if it would be able to lure and demonize little children and youth into sexual feelings once it would be digested. The cookie indeed has a very alluring, friendly, and delicious appearance, and became associated menace with the gender phantasm. Previously, in September 2021, Geert Wilders, who was then a member of the Dutch Tweede Kamer (House of Representatives), debated LTBQIA+ proponent Rob Jetten of D66, and gave a very emotional speech about the irritating excesses of left-wing society that would impinge on a "normal life"—including diversity and LGBTQIA+ rights, pro-migration politics, EU treaties, climate activism and the dictatorship of "woke" Dutch culture. He labeled all of these as a "national suicide letter" and compared these initiatives to totalitarian regimes that forbid specific rights and freedoms (@Arnews, 2023). His tirade against Dutch liberal culture included a harrowing outcry about secondary school students who were being inundated with genderkoek.

The political leader of Vlaams Belang Tom Van Grieken further amplified these ideas in 2024 and made attacks on gender ideology as a primary election point. In a televised debate about "woke" culture he confronted transgender

Green Party frontrunner and vice premier, Petra de Sutter, when he said that he would only recognize her as a "man" (DeMorgen 2024). In June 2024, Vlaams Belang also started attacking the more moderate nationalist party N-VA for funding genderkoek education and for supporting LGBTQIA+ rights. In the Federal elections in both the Netherlands and Belgium, which took place in November 2023 and June 2024 respectively, the far-right campaigns included further antics over genderkoek. On platforms X and TikTok, various individuals voiced their disgust over genderkoek. A promotional video by Vlaams Belang issued during the 2024 elections combines and image of the "wielder of death" with a hysterical voice narrating that the LGBTQIA+ organization Çavaria receives public subsidies to introduce genderkoek to children. The voice is set onto industrial-experimental rowdy music and says the children should be left alone (@Vlaams Belang, May 11, 2023). Genderkoek was discussed on X by many Flemish and Dutch citizens who saw it as a "toxic indoctrination," or representative of "preachy, irritating sexual ethics" and they want to be "left alone." One citizen commented that people can do whatever they want in their bedrooms, but there should be no public policies about that (@Marchionatus, June 4, 2024). Another citizen posited that genderkoek was a threat to society and would coincide with ethnic minorities becoming dominant in schools and beating up white children (@Loewiedefinesse, June 9, 2024). Digestion of this cookie would make their child delirious and "nauseous" (@Equilibrium, June 7, 2024). Homosexuality as such is not a problem amongst white Belgians, but it was a problem that genderkoek was aggressively "pushed down the throats" of innocent children (@viv, June 6, 2026). As a matter of fact, the cookie and the entire field of Gender Studies was rejected as "invalid science" (@Thoma Spaas, June 4, 2024). The cookie was also related to the presence of LGBTQIA+ flags in Belgian workspaces, which is deemed as another type of indoctrination by "half-humans"(@ Mittemeijer, July 5 2024). These two "infantile" tools of gender ideology evoked horror, revulsion, expressions of violence, and once again an open declaration that these tools are also used by pedophiles to prey on children. Similar expressions on TikTok were turned into full-blown and well edited music videos. In a supposedly snappy re-edit of Wilder's 2021 speech, issued in June 2024 when Wilders paid a visit to Belgium to support VB, the speech was cut down and set onto industrial-melancholic techno beats (@Chazie, 9 May 2024).

These attacks on gender were then taken up by the Flemish national news (VRT) as part of their campaigns to promote voting amongst youth. Young adults aged sixteen and above were invited to dedicated TV programs to debate wokeness and other "hot topics" with various politicians. In a TV show for first-time voters (called Eerste Keus), VB politician Chris Janssens, who is openly gay, debated gender ideology with Conner Rousseau, who came out as bisexual, of the socialist party Vooruit. One young adult asked him about sex and gender education, and he answered ""I am a homosexual, but I am not a proponent of gender ideology. In the classroom one should concern oneself with ABCDE and not LGBTQIA+"

(Decré 2024). He also reiterated the exact lines of Geert Wilders saying "[…] that children need to be left alone at school" and that when their hormonal levels change "[…] they will know what to do with their sexual orientation." The young man asked him at what age such lessons would be relevant at school and he replied, "at no age is this relevant," at which point many in the audience cheered and applauded in agreement. Janssens was then confronted by Conner Rousseau, who defended LGBTQIA+ identities and stated that people should be able to live with these identities. Another youth was not swayed and kept asking Conner about a 4-year-old girl, who would be told that "[…] It would be ok to have relations with another 4-year-old girl." (VRT 2024) It is strange that the conversation veered into discussions of children's sexuality, even though it was meant to be about first-time voters in Belgium, who are aged sixteen or above. The fact that young adults were invited to discuss gender and sexuality without having any background information shows how the news media zoomed in on "the problem" of youth sexualization. Even though these ideas were largely based on phantasms and conspiracy theories, they were reinvigorated by far-right political parties and taken up by a mainstream news channel.

Conclusion

At the time of finishing the essay in Spring 2025, the election of Trump once again as president of the USA has given an enormous impetus to USA attacks on feminism and sexual minorities and has invigorated global anti-gender movements. This article gives evidence about the fact that the attacks on DEI (Diversity, Equality, and Inclusion) policies and sex-gender education and politics were also alarmingly taken over public culture and the news media within the EU. We focused on online discourses on X of far-right politicians and grassroots organization both in Wallonia and Flanders, with a focus on how they are driven by "phantasms" or conspiratorial fears about the education and sexualization of children. A fear of sexualization was packed together with many other social fears, such as the fear of a dwindling nation and a fear of loss of strictly defined binary gender roles. It became almost literally the case that democratic and left-leaning policies became associated with sexual misconduct towards innocent victims, our children.

These ideas were driven by far-right politicians and heterogeneous actors who packaged their different political visions as a condensed and coherent emotional outcry about the protection of children. It is also the case that youth in Belgium and the EU are quickly influenced and transformed by the antics of these anti-gender movements. As shown in one of the 2024 elections programs for youth on VRT, young men stated out of the blue that they were afraid that their younger sister or brother would be approached and seduced by pedophiles (VRT 2024). In Spring 2025, the results of the Gender Equality Index of 2024 were published, showing a systematic bias of Belgian youth towards gender. Although Belgium supposedly holds a place in the "top five gender-equal countries in Europe," 50 percent of young Belgians responded that it is acceptable for a man to control his wife's/spouse's financial means.(VRT 2025) Similarly, in another survey that was published several months later, it was observed that Belgian youth is also becoming far less tolerant towards LGBTQIA+ rights and identities (Eelbode 2025). At the same time, there is reason to remain hopeful. As reported in the Belgian media, there are also several political parties and activists, as well as many ordinary citizens who are not swayed by the radical hysterias and conspiracies of these anti-gender movements. (Santens 2024). People from different backgrounds, from grandmas in their 70s to students in their 20s, testify about the importance of sexual education and dismantle the radical conservatism and conspiracies of the far-right.

References

Arnews. 2021. "'Doe eens normaal zeg' v Rob Jetten - Algemene Politieke Beschouwingen Tweede Kamer" YouTube, 22 September 2021 Accessed 7 July 2025 at: https://www.youtube.com/watch?app=desktop&v=8o6NQnXIPkM&fbclid=IwZXh0bgNhZW0CMTAAAR3d6K5WtLLEPjHBA61Twmeuw8IyPE5Y2qZZvmsbEVh0u9i_08yf_lis0A0_aem_AZBl-sl-oy6VZT66F3IeEfPEEYrGigGuWFHhN2LcyeVZh9bsIQIYJuAUZLnJqKwZ-fdgc_Wi_18HDCYR9HEwxd46

Baumers, Koen, and Kim Clemens. 2023. "Het begon met open brief van kinderpsychiater, maar toen ontstonden de meest absurde theorieën over lessen seksuele opvoeding." *Het Nieuwsblad*, September 15, 2023. https://www.nieuwsblad.be/cnt/dmf20230914_96887619.

Bezorgde Ouders, 2023 Accessed 21 March at https://www.bezorgdeouders.be/wie-zijn-wij 2024); The webinar against gender ideology primary and secondary education was accessed March 22, 2025 at https://www.youtube.com/watch?v=yVFbhR7yPBs

Bon Sens Belgique. n.d. "A PROPOS." Accessed December 10, 2023. https://bonsensbelgique.be/a-propos/.

Butler, J. 2024. *Who's Afraid of Gender*. New York: Penguin Books.

Bossner, Felix, and Melanie Nagel. 2020. "Discourse Networks and Dual Screening: Analyzing Roles, Content and Motivations in Political Twitter Conversations." *Politics and Governance* 8 (2): 311–25. https://doi.org/10.17645/pag.v8i2.2573.

Butler, J. and Syed T. 2024. "Transnational anti-gender politics and resistance." London School of Economics Lecture, Accessed 20 March 2025, https://www.youtube.com/watch?v=AmLiW_tuyy0

BRUZZ. 2023a. "200 Deelnemers Bij Nieuwe Betoging Tegen Evras-decreet." *Bruzz*. October 1, 2023. Accessed December 8, 2023. https://www.bruzz.be/samenleving/200-deelnemers-bij-nieuwe-betoging-tegen-evras-decreet-2023-10-01.

———. 2023b. "Anti-Evras Betoging Aan Kunstberg Trekt 150 Manifestanten." October 15, 2023. Accessed December 9, 2023. https://www.bruzz.be/onderwijs/opnieuw-anti-evras-betoging-aan-de-kunstberg-2023-10-15.

Chazie, Geert Wilders' remix, TikTok, May 9 2024, Accessed 7 July 2024 athttps://www.tiktok.com/@chazie.ae/video/7366952162301087009?_r=1&_t=8mv9sxqw4vo

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. "The Echo Chamber Effect on Social Media." *Proceedings of the National Academy of Sciences* 118 (9). https://doi.org/10.1073/pnas.2023301118.

De Wolff, Daniel [@DanieldeWolff]. 2023. "#EVRAS @alexanderdecroo @CarolineDesir #Belgique #OMS #Agenda2030 #GreatReset." X. Accessed April 30, 2024. https://twitter.com/DanieldeWolff/status/1740473687722033391.

Datta, N. 2018. *Restoring the Natural Order." EU Parliamentary Forum for Sexual & Reproductive Rights Report*, Accessed 18 May 2024 https://www.epfweb.org/node/175

Debruyne, Heleen. 2023. ""'Ze zeggen aan kleuters dat geslachtsoperaties leuk zijn": wie het Evras-programma zo interpreteert, heeft toch een rare kronkel.'" *Humo*. September 20, 2023. Accessed December 9, 2023. https://www.humo.be/meningen/ze-zeggen-aan-kleuters-dat-geslachtsoperaties-leuk-zijn-wie-het-evras-programma-zo-interpreteert-heeft-toch-een-rare-kronkel~b6069d07f/.

De Coninck, Douglas. 2023. "Betoging in Brussel tegen lessen seksuele opvoeding: 'Dit is een front van het geweten.'" *De Morgen*, September 17, 2023. https://www.demorgen.be/nieuws/betoging-in-brussel-tegen-lessen-seksuele-opvoeding-dit-is-een-front-van-het-geweten~b4dc94cd/.

Decré, Hanne. 2023. "Brandstichting in Vier Scholen in Charleroi En Alles Lijkt Te Draaien Rond 'Evras', Het Nieuwe Decreet Over Seksuele Opvoeding." *VRT NWS,* September 13, 2023. https://www.vrt.be/vrtnws/nl/2023/09/13/brandstichting-scholen-charleroi-protest-lessen-seksuele-opvoedi/.

Decré, H. 2024. "Moet er gesproken worden over Geaardheid in de Klas" *VRTnieuws,* 3 June. Accessed 7 July 2024 at : https://www.vrt.be/vrtnws/nl/2024/06/02/eerste-keus-verkiezingen-jongerengeaardheid/#:~:text=Rousseau%3A%20%22We%20gaan%20ons%20onderwijs,vraag%20me%20af%20waarom%20niet.%22

De Lobel, Peter. 2023. “‘Neen, We Gaan Kinderen Geen Gender Opdringen of Leren Masturberen.’” *De Standaard*, September 14, 2023. https://www.standaard.be/cnt/dmf20230914_97245988.

DeMorgen 2024. “Vlaams Belang-voorzitter Tom Van Grieken heeft ‘excuses aangeboden’ aan Groen-vicepremier Petra De Sutter”, 19 June, Accessed 7 July 2024 at https://www.demorgen.be/snelnieuws/vlaams-belang-voorzitter-tom-van-grieken-heeft-excuses-aangeboden-aan-groen-vicepremier-petra-de-sutter~befb4f44a/?referrer=https://www.google.com/

Edenborg, Emil. 2021. “Anti-Gender Politics as Discourse Coalitions: Russia’s Domestic and International Promotion of ‘Traditional Values.’” *Problems of Post-Communism* 70 (2): 175–84. https://doi.org/10.1080/10758216.2021.1987269.

Eelbode, Frederic. 2025, “Jeugd is minder tolerant voor holebi’s. Progressive trend is gekeerd.” *De Standaard*, 3 April, Accessed 8 June 2025 at https://www.standaard.be/binnenland/jeugd-is-minder-tolerant-voor-holebi-s-progressieve-trend-is-gekeerd/52867509.html.

Equilibrium, Gendercookie makes one nauseuos, 9 June 2024, X, Accessed 7 July 2024 at https://x.com/SuskeSambelvi/status/1799162833336258611?s=07

Evans, Michael P., Andrew Saultz, and Sue Winton. 2021. “Social Media Utilization in Discourse Coalitions: The Opt-Out Movement in Ohio.” *Teachers College Record the Voice of Scholarship in Education* 123 (5): 1–26. https://doi.org/10.1177/016146812112300509.

https://x.com/SuskeSambelvi/status/1799157122770542952?s=07&fbclid=IwZXh0bgNhZW0CMTAAAR03W8RmIoJRSrPyveRdRA4kqPv90nQG2KxZzTcB1CdB3BNlRUNQzOsiaR4_aem_AZBoDZViDyKn7mwuj6BzpktS031eB6O6ZreafFoMfaPAPlg8liH9XZ-T6iHMdYMG4SmO7y68OJ1hRm31G7qorFTI

EVRAS. n.d. “L’EVRAS À L’école - EVRAS - Parents.” Accessed December 2, 2023a. https://www.parent.evras.be/levras-a-lecole/.

———. n.d. “Une mission obligatoire.” Accessed December 2, 2023b. https://www.evras.be/evras-bien-plus-que-leducation-sexuelle/cest-quoi-levras/une-mission-obligatoire/.

Gillani, Nabeel, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. “Me, My Echo Chamber, and I.” WWW ’18: *The Web Conference 2018*. Lyon, France, January 1. https://doi.org/10.1145/3178876.3186130.

Graff, Agnieszka, and Elżbieta Korolczuk. 2021. *Anti-Gender Politics in the Populist Moment*. https://doi.org/10.4324/9781003133520.

Hajer, M.A. 2006. “Doing Discourse Analysis: Coalitions, Practices, Meaning." *Nederlandse Geografische Studies*, 65–74. https://dare.uva.nl/personal/pure/en/publications/doing-discourse-analysis-coalitions-practices-meaning(e9a062b7-099f-4f6c-b4e2-a018a0170da5).html.

Justaert, Marjan. 2023. "Grimmig Protest: 2.000 Demonstranten Eisen Intrekking 'Pedofiliedecreet.'" *De Standaard*, September 17, 2023. https://www.standaard.be/cnt/dmf20230917_95587075.

Kairos. n.d. "A Propos De Kairos." Accessed February 17, 2024. https://www.kairospresse.be/a-propos-de-kairos/.

Kuhar R. and D. Paternotte. 2017. *Anti-Gender Campaigns in Europe: Mobilizing Against Equality,* Maryland: Rowman & Littlefield.

La Zone Libre. n.d. "Bienvenue." Accessed December 10, 2023. https://www.free-zone-movement.net/fr/.

La Zone Libre - Belgique [@LaZoneLibre]. 2023. "Ce soir, 19:00 Auderghem ! #Agenda2030 #thegreatresist. #WHO #OMS @WHO @WHOatEU." X. Accessed March 11, 2024. https://twitter.com/LaZoneLibre/status/1727838443089047986.

Loewiedefinesse, gendercookie is a threat to society, June 9 2024, X, Accessed 7 July 2024 at https://x.com/loewiedefinesse/status/1799712232743514515?t=R6VbkPUUmVnK2BzljD_t1Q&s=07&fbclid=IwZXh0bgNhZW0CMTAAAR07ku4N5KIbQCMllB6rSQ08haQpCLua6jSNmmfc_slGWnY0t4sYifKNqP0_aem_AZBG91cNYfIelUJeIyQw6tQssvaYZB4bueLwtdnjSjMVjj6E63nfBHG5uPVD6Y0F1e-7t6iCMnhTYxgRW7edwBmi

Marchionatus, people can do whatever in their bedroom June 4 2024. X, Accessed 7 July 2024 at https://x.com/marchionatus/status/1797900438345507112?t=PZV1KVtbalpc2QdCESMG_w&s=07&fbclid=IwZXh0bgNhZW0CMTAAAR03W8RmIoJRSrPyveRdRA4kqPv90nQG2KxZzTcB1CdB3BNlRUNQzOsiaR4_aem_AZBoDZViDyKn7mwuj6BzpktS031eB6O6ZreafFoMfaPAPlg8liH9XZ-T6iHMdYMG4SmO7y68OJ1hRm31G7qorFTI

Mittemeijer, Friend of Mine had an inclusivity training, July 5 2024, X, Accessed 21 March 2024 at https://x.com/MittemeijerOote/status/1809128986577924342?s=07&fbclid=IwZXh0bgNhZW0CMTAAAR12LsoDeBGW-ly8svOXxxL50ZvK8OKYfwxIoilWs_v_tNOfbdDDMv1vgjM_aem_vEzIt7vw3APpFZtAN9Y_6w (Accessed 7 July, 2024)

Mouhamou, Inass. 2023. "Islamitische Organisaties Stappen Naar Grondwettelijk Hof Tegen Lessenpakket Evras Over Seks En Relaties." *vrtnws.be*, October 2, 2023. https://www.vrt.be/vrtnws/nl/2023/10/02/meerdere-islamitische-organisaties-naar-grondwettelijk-hof-tegen/.

Norris, S. 2023. *Bodies Under Siege: How the Far-Right Attack on Reproductive*

*Rights went Globa*l, New York Verso Books.

North, Anna. 2020. "How #SaveTheChildren Is Pulling American Moms Into QAnon." *Vox*, September 18, 2020. https://www.vox.com/21436671/save-our-children-hashtag-qanon-pizzagate.

Paternotte, David, and Roman Kuhar. 2018. "Disentangling and Locating the 'Global Right': Anti-Gender Campaigns in Europe." *Politics and Governance* 6 (3): 6–19. https://doi.org/10.17645/pag.v6i3.1557.

Pontifical Council for the Laity. 2011, "Interview with Marguerite Peeters on the Gender Theory," September 26, Accessed 21 March 2025 at http://www.laici.va/content/laici/en/sezioni/donna/notizie/interview-with-marguerite-a--peeters-on-the-gender-theory.html?fbclid=IwAR2HkQ-yxvcs9LsmqEk6RhFQRWhYOT-pQkcgu4FWOdljMs-aZ0WhaOnyOvQo

Redactie De Morgen. 2023. "Meerdere islamitische organisaties trekken naar Grondwettelijk Hof tegen Evras." *De Morgen*, October 2, 2023. https://www.demorgen.be/snelnieuws/meerdere-islamitische-organisaties-trekken-naar-grondwettelijk-hof-tegen-evras~b806fe48/.

Rédaction RTBF. 2023. "Complotistes, Extrême Droite Et Adeptes De Théories Pédocriminelles : Voici Le Réseau Des Désinformateurs Sur L'Evras En Belgique - *RTBF Actus*." RTBF. September 28, 2023. Accessed February 17, 2024. https://www.rtbf.be/article/complotistes-extreme-droite-et-adeptes-de-theories-pedocriminelles-voici-le-reseau-des-desinformateurs-sur-levras-en-belgique-11256548.

ROHFF [@rohff]. 2023. "Si vous refusez qu'on apprenne à vos enfants,vos neveux et nièces de 5 ans à se masturber et qu'on." X. Accessed March 10, 2024. https://x.com/rohff/status/1702069273399599192.

Santens, Tobias. 2024. "Tom Van Grieken: 'Vlaams Belang Gelooft Niet Dat Er Genders Zijn', Kijker Lucienne (77): "Ik Ben 2 Keer Zo Oud Als U, Maar Voel Me Moderner" | *VRT NWS*: Nieuws." VRT NWS, May 26, 2024. https://www.vrt.be/vrtnws/nl/2024/05/26/vlaams-belang-gender-lucienne-de-zevende-dag/.

Sauvons Nos Enfants. n.d. "Accueil." Accessed December 10, 2023a. https://sauvonsnosenfants.weebly.com/.

———. n.d. "Lois." Accessed December 10, 2023b. https://sauvonsnosenfants.weebly.com/lois.html.

Stroobants. J.P .2023. "In Belgium, several schools set on fire after extremist campaign against sex education, *Le Monde*. Sept 18, Accessed 21 March 2025. At https://www.lemonde.fr/en/international/article/2023/09/18/in-belgium-several-schools-set-on-fire-after-extremist-campaign-against-sex-education_6137195_4.html

Struys, Bruno. 2023. “Katholieken, moslims en uiterst rechts: achter het protest tegen lessen seksuele opvoeding zit een vreemde alliantie.” *De Morgen*, September 20, 2023. https://www.demorgen.be/nieuws/katholieken-moslims-en-uiterst-rechts-achter-het-protest-tegen-lessen-seksuele-opvoeding-zit-een-vreemde-alliantie~ba4cc281/.

Terren, Ludovic, and Rosa Borge-Bravo. 2021. “Echo Chambers on Social Media: A Systematic Review of the Literature.” *Review of Communication Research* 9 (January): 99–118. https://doi.org/10.12840/issn.2255-4165.028.

Titeca, Kristof. 2023. “Protest Tegen Evras: Onverwachte Coalities Op De Barricaden.” *De Standaard*, September 19, 2023. https://www.standaard.be/cnt/dmf20230918_96300863.

Thomas Spaas, Gendercookie and science, X, June 4 2024, Accessed 7 July 2025 at https://x.com/thomas_spaas/status/1798073516866273461?s=07&fbclid=IwZXh0bgNhZW0CMTAAAR1tUv-hOrJ8zlUCdYOFOxeUXmQfDJQESb51F4JIMgDHEVFETBBa5krquXU_aem_AZBUNL1dh9Ys_-.MAEEfgPFzhX67o1O1alY_3ipG64KZHUIoJE8U46qVY_2HhvxH5c9iK5guldntATcD13N1D0v7C

Unesco Platform Vlaanderen. 2018. “Verenigde Naties Pleiten Voor Uitgebreide Seksuele Opvoeding.” *Unesco Vlaanderen*. January 24, 2018. Accessed December 8, 2023. https://www.unesco-vlaanderen.be/unesco-in-de-kijker/nieuws/verenigde-naties-pleiten-voor-uitgebreide-seksuele-opvoeding.

Van Bakel, Luc, Ellen Debackere, and Amra Dorjbayar. 2023. “Nepnieuws Linkt LGBTQ+-gemeenschap Steeds Vaker Aan Geweld En Pedofilie | VRT NWS: Nieuws.” VRT NWS, July 4, 2023. https://www.vrt.be/vrtnws/nl/2023/06/28/fysiek-geweld-op-lgbtq-personen-blijft-hoog-nepnieuws-speelt-b/.

Vancaeneghem, Jens. 2023. “Horrorfilmpjes, Open Brieven En Brand Op Kleuterscholen: Hoe Controverse Over 2 Uur Seksuele Opvoeding per Jaar Ontspoorde.” *Gazet Van Antwerpen*, September 13, 2023. https://www.gva.be/cnt/dmf20230913_97112697.

Van Maele, Pieter, and Peter De Lobel. 2023. “Van Conservatieve Moslims Tot Kinderpsychiater: Ongewone Coalitie Stookt Protest Tegen Seksuele Opvoeding Op.” *De Standaard*, September 15, Accessed June 8 2025 at https://www.standaard.be/binnenland/van-conservatieve-moslims-tot-kinderpsychiater-ongewone-coalitie-stookt-protest-tegen-seksuele-opvoeding-op/40738216.html.

Vlaams Belang, 2024, Eerst keus Verkiezingen, TikTok, Accessed 21 March 2025 at https://www.tiktok.com/@vlaams_belang6/photo/7366754785430687008?_d=secCgYIASAHKAESPgo8KnVvnEJ2zzQUCqKU6mmwx0IBWpgLL%2FtCTkI4x8VlFJjcuYsodCt8OgG7BO92SxluRp1UISYZgGId%2B5izGgA%3D&_r=1&aweme_type=150&checksum=94fb474bbdb89316ceb778

a728c6d6b87fb4e9315afc529bb5ba9c8fe9d5d17e&pic_cnt=5&preview_pb=0&sec_user_id=MS4wLjABAAAA0MxxBF7UGP4yU1UCYLsavysarA9j-G4cIaePz4p5KGUcieR-BsCuIC6Lx_RUxEws&share_app_id=1233&share_item_id=7366754785430687008&share_link_id=8d2ee7bf-6510-4230-96ac-21d83fb9dadf&sharer_language=en&social_share_type=0&source=h5_m&timestamp=1717942329&u_code=eedmlke653746c&ug_btm=b2001%2Cb7464&ug_photo_idx=0&ugbiz_name=UNKNOWN&user_id=7376651158272885793&utm_campaign=client_share&utm_medium=android&utm_source=messenger

Viv, Pushed down the throats of innocent children, June 6 2024, X, Accessed 7 July 2024, at https://x.com/Viv40351916/status/1798546687591665752?s=07

VRT 2023a, Tom Van Grieken, Isolde Van Den Eynde, and Stijn Baert. 2023. "Podcast 'De Afspraak Op Vrijdag' Met Tom Van Grieken, Isolde Van Den Eynde En Stijn Baert." December 10, December 10, Accessed 8 June 2025 at https://www.vrt.be/vrtnws/nl/2023/09/15/podcast-de-afspraak-op-vrijdag-met-tom-van-grieken-isolde-van/.

VRT 2023b. "Nieuw Protest Tegen Verplichte Lessen Seksuele Opvoeding in Franstalig Onderwijs, Ministers Houden Voet Bij Stuk.", September 17, Accessed 6 June 2025 at https://www.vrt.be/vrtnws/nl/2023/09/17/zo-n-200-mensen-protesteren-in-brussel-tegen-verplichte-lessen-s/.

VRT 2024, "Eerst Keus Verkiezingen," June 3, Accessed 7 July 2024. https://www.vrt.be/vrtnws/nl/2024/06/02/eerste-keus-verkiezingen-jongeren-geaardheid/#:~:text=Rousseau%3A%20%22We%20gaan%20ons%20onderwijs,vraag%20me%20af%20waarom%20niet.%22

VRT 2025, "Steeds meer jonde Europese mannen hebben vrouwonvriendelijke ideeën," March 7 2024, Accessed 5 June 2025 at https://www.vrt.be/vrtnws/nl/2025/03/07/steeds-meer-jonge-europese-mannen-hebben-vrouwonvriendelijke-ide

Part IV

Media, Journalism, and Countering Hate Narrative

Beyond Hate: Disinformation Targeting Minorities and the Erosion of Democratic Debate

Carla Sentí Navarro, Human Rights Institute of the University of Valencia

Introduction

Public debates in contemporary democracies increasingly take place in digital environments, where information travels at a speed and scale that were unthinkable only a few decades ago. These spaces, shaped by algorithmic visibility and high levels of user engagement, have become fertile ground not only for the exchange of ideas, but also for the rapid dissemination of misleading or manipulative content. Within this broader information disorder, one phenomenon has gained worrying prominence: the spread of prejudiced narratives targeting vulnerable and minority groups.

Social media platforms, in particular, allow false or distorted claims to spread widely and with unprecedented speed, often framing migrant communities, ethnic or religious minorities, or LGTBQ+ individuals as social threats. This wave of prejudiced disinformation is, however, not an isolated phenomenon but part of a broader transformation in the way public opinion is shaped and political agendas are built in contemporary democracies.

The relevance of this problem extends far beyond the online sphere. The European Parliament has recognized the weaponization of disinformation against minorities (Szakács, J. and Bognár, É., 2021). Socially, these narratives reinforce stigma and deepen existing inequalities; politically, they polarize public debate and weaken democratic institutions; legally, they change the capacity of current regulatory frameworks to balance freedom of expression with the need to safeguard dignity and equality. While many European jurisdictions, including Spain, criminalize certain forms of hate speech, the harms produced by prejudiced disinformation exceed the scope of individual offences. They affect not only the dignity and safety of targeted groups but also the health of the democratic information environment.

From this perspective, it is useful to distinguish between two intertwined forms of harm. The first is a direct harm: the tangible and symbolic consequences suffered by vulnerable communities, such as increased hostility, discrimination, or violence towards them. The second is a structural harm, which affects democratic societies as a whole. When false narratives gain popularity, they distort the public debate, erode trust in institutions, and limit citizen's ability to make informed decisions. These effects are especially visible during periods of political uncertainty, tension, or electoral competition, when manipulative content tends to escalate.

The aim of this article is to examine how prejudiced disinformation operates at both levels and to explore the implications this dual harm has for democratic resilience. Three main questions are addressed: How do these narratives take shape and circulate in digital environments? What legal frameworks exist to address them? And what kind of responses might better capture the complexity of the harm involved?

2. Conceptual framework

In order to understand the double problematic that these discourses have at an individual and collective level, we must clarify the conceptual foundations that underpin prejudiced disinformation and the dynamics that allow it to circulate. This requires distinguishing it from broader forms of information disorder, situating it within longer histories of exclusion affecting minorities and vulnerable groups, and examining how platform architectures shape the visibility and persuasive power of such narratives.

1. Prejudiced disinformation

The concept of prejudiced disinformation refers to the deliberate creation or the dissemination of false, misleading, or distorted claims that portray minority groups as dangerous or socially corrosive. It combines two dimensions: intentional informational manipulation and the targeting of communities characterized by histories of discrimination and marginalization. As it holds this dual nature, prejudiced disinformation cannot be neatly under existing labels within the information disorder literature.

Although related, it differs from misinformation – false content shared without intention to deceive – and malinformation, which involves the selective use of truthful information in misleading contexts (Wardle, C. and Derakshan, H., 2017). It also cannot be fully assimilated to propaganda, traditionally understood as state-driven or politically orchestrated persuasion (Freund, J., 1968), nor it is simply reducible to hate speech. While prejudiced disinformation is prone to foster environments that facilitate hostility or violence, and hate towards certain groups, it often does so through indirect mechanisms: insinuation, suggestive framing, repetition of stereotypes, or the amplification of fabricated events. In many cases, the messages appear plausible or are presented as concerns, questions or warnings, which complicates their classification under legal categories that require explicit incitement.

This distinction is analytically important. Many forms of prejudiced disinformation fall outside existing regulatory frameworks in jurisdictions like Spain, where criminal law focuses on expressions of hatred or hostility, not on the manipulation of factual claims that may indirectly fuel discriminatory attitudes. Understanding prejudiced disinformation requires a conceptual approach attentive not only to the veracity of statements, but also to the social meanings and impacts they produce and political work they perform.

2. Minority and vulnerable groups

We understand minority groups as a "group of persons which constitute less than half of the population whose members share common characteristics of culture, religion, language, [sexual orientation or ethnicity], or a combination of any of these" (Office of the United Nations High Commissioner for Human Rights [OHCHR], n.d.). This notion includes both formally recognised minorities and socially stigmatised collectives that suffer systemic exclusion. Vulnerability, in this context, is not an inherent trait, but a position within structures of social, economic or political inequality that increases exposure to harm (UN Special Rapporteur on Minority Issues, 2019).

Prejudiced disinformation plays on existing stereotypes and historicized forms of "othering" to activate fear, disgust or moral panic. For instance, migrants may be portrayed as criminals or threats to welfare systems (Arcila Calderón, C. et. al., 2022); Muslims as inherently violent (Fuentes Lara, C. and Arcila Calderón, C., 2023); or LGBTQ+ individuals as undermining family values (Strand, C. and Svensson, J., 2021). These narratives do not emerge in a vacuum, instead, they draw on historically sedimented regimes of knowledge and power that reproduce exclusionary understandings of the "other", as described in Foucault's notion of historically embedded discursive formations (1991) and in Gilroy's analysis of the enduring legacies of colonial racial imaginaries (2004).Their harmfulness lies not only in being false or misleading, but in their resonance with discriminatory logics that remain operative in contemporary societies.

3. Information Disorders and the Digital Public Sphere

The proliferation of prejudiced disinformation must also be understood in relation to the structure of the digital environments in which it circulates. Social media platforms operate through engagement-based visibility: content that generates strong emotional reactions, such as fear, anger or outrage is more likely to be prioritized by algorithms (Arias Maldonado, M., 2016). Prejudiced narratives, which often activate powerful affective responses, benefit disproportionately from this system of amplification.

Digital spaces are hybrid arenas where political actors, media outlets, influencers, and ordinary users coexist. The boundaries between professional journalism, political communication, and user-generated content become increasingly blurred. This hybridity allows narratives to circulate across different communities, platforms and formats (Chadwick, A., 2017). The speed and scale of this process make corrective efforts difficult, allowing falsehoods to shape public perceptions before reliable information gains visibility.

The broader concept of information disorders captures these dynamics. Today's information environment is not only characterized by the presence of falsehoods but also by fragmentation, overload, and strategic manipulation (Wardle, C. and Derakshan, H., 2017). Prejudiced disinformation thrives in this context by offering emotionally compelling explanations that resonate with identity-based grievances. Its spread is facilitated not only by the intentional actions of political entrepreneurs but also by platform architectures, recommender systems, and economic incentives that reward attention-grabbing content over accuracy.

At the same time, the digital public sphere challenges traditional gatekeeping roles. Whereas mainstream media once filtered and contextualized information, digital platforms decentralize communicative authority (Benkler, Y., Faris, R., and Roberts, H., 2018). This democratization expands participation but also increases vulnerability: coordinated networks can artificially inflate the visibility of certain claims; automated accounts can amplify narratives at scale; and monetization models encourage sensationalism.

Recognizing these structural conditions is very important. Prejudiced disinformation is not merely a collection of false statements but a phenomenon embedded in the logics of the contemporary information ecosystem. Its harms, both direct and structural, arise from the interaction between platform design, social biases, and political incentives. The following sections examine these harms in detail.

3. Direct and Structural harm

Direct harm

Prejudiced disinformation – false or misleading content targeting people based on their ethnicity, religion, sexual orientation or other identity – has tangible and pernicious effects on those communities.

One immediate harm is the increase in discriminatory attitudes and acts. Information manipulation campaigns have been shown to "contribute to increasing hatred against minorities", producing a "direct negative impact on the fundamental right to human dignity" (Szakács, J. and Bognár, É., 2021).

More alarmingly, prejudiced falsehoods can incite hate crimes and violence. Social media provides a frictionless conduit for demonizing narratives to spread and metastasize into offline action. Empirical research is increasingly drawing a link between online hate and physical attacks: surges in Islamophobic or anti-migrant memes and slogans on platforms often anticipate (but do not necessarily causally determine) subsequent spikes in violent hate crimes. For instance, a 2024 study of Spain found that fluctuations in anti-immigrant and anti-LGBT hate speech on Twitter and Facebook correlated with rises in hate crime reports, suggesting that online inflammatory language could be a leading indicator for offline violence (Arcila Calderón, C. et. al., 2024). The causal pathways are complex, but the trend is clear: when a group is incessantly vilified as a threatening "other", whether Roma, Muslim, Jewish, Black, or LGBTQ, some individuals eventually act on those demonizing narratives, with vandalism, assaults or worse.

Beyond violence by private actors, prejudiced disinformation can also drive policy exclusion and institutional maltreatment of minority groups. False narratives create public support for harsh measures that target those depicted as dangerous or unworthy. During the first wave of COVID-19, for example, conspiracy theories alleging that migrants were covertly spreading the virus gained traction in several countries. These unfounded claims were not benign: in Hungary, they were invoked by the government to justify suspending asylum admissions for refugees in March 2020, purportedly to protect public health (Montalto Monella, L. and Palfi, R., 2020). In Italy and Spain, similar rumors about "infected migrants…escaping quarantine or purportedly infecting police officers" circulated widely (The Poynter Institute), bolstering calls for tough anti-immigration policies. Disinformation thus becomes a pretext for what amounts to institutional discrimination – laws or practices that exclude or overly scrutinize marginalized groups. Even in less extreme cases, prejudicial falsehoods have a "chilling effect" on civil society support for minorities. Advocacy NGOs may withdraw for vilified groups, fearing loss of credibility by association (Szakács, J. and Bognár, É., 2021). In this way, disinformation-fueled hatred doesn't just intimidate the vulnerable group itself; it also isolates them from the solidarity of allies. All these factors – from social alienation and hate-fueled violence to exclusionary public policies – underscore that prejudiced disinformation inflicts direct, multidimensional harm on vulnerable communities, threatening not only their safety but their equal standing in society.

Structural harm

In addition to the harm it causes to targeted minorities, prejudiced disinformation inflicts structural harm on democracy itself. Liberal democracy depends on a well-informed citizenship, a pluralistic public sphere, and trust in institutions (Dahl, R., 1956). Each of these pillars is weakened when propaganda and falsehoods distort public discourse. Unlike legitimate debate, which in democratic societies protects even views that may "offend, shock or disturb" (ECtHR, Handyside v. United Kingdom, 1976), disinformation does not merely introduce a contentious perspective. As the Court has repeatedly clarified, freedom of expression does not extend to demonstrably false factual allegations (ECtHR, Lingens v. Austria, 1986), nor does it exempt communicators from the responsibility to contextualize harmful claims (ECtHR, Jersild v. Denmark, 1994). In the digital sphere, moreover, platforms may contribute to the amplification of such narratives at scale (ECtHR, Delfi AS v. Estonia, 2015).

When significant segments of the population believe, for example, that immigrants are systematically dangerous, that minority faiths are plotting against society, or other such baseless claims, democratic dialogue shifts away from reality-based problem-solving. This shift crowds out meaningful discussion: political attention is redirected towards invented dangers while actual social challenges remain overlooked. In other words, the "marketplace of ideas" becomes polluted, making it more difficult for democratic decision-making to rest on a shared understanding of facts.

Moreover, prejudiced disinformation corrodes the social cohesion and trust that underpin a healthy democracy. By targeting a social group and spreading distorted negative information about it, such campaigns reinforce the notion of an "out-group" in society. They undermine intercommunity solidarity. The goal – and effect – is to drive a wedge between communities, eroding citizens' confidence that state institutions will treat all groups fairly.

This mutual mistrust is negative for democratic pluralism. The European Parliament's Committee on Foreign Interference noted that such tactics can even "undermine social cohesion" and diminish respect for the rule of law. Indeed, when prejudiced falsehoods proliferate, citizens may start doubting the integrity of information coming from official sources or mainstream media, a doubt fueled by extremist voices. Over time, this can translate into a broader erosion of institutional trust, where people no longer know who or what to believe. Democratic institutions – from elections to courts to the press – rely on public confidence; disinformation actively works to shatter that confidence (World Economic Forum, 2017), often by painting authorities as complicit in covering up "the truth" about the targeted minority, a common trope in conspiracy theories (Taylor, A., 2016).

The problem is aggravated by the way digital platforms work. Algorithms designed to maximise engagement tend to give more visibility to sensational and emotionally charged content, which generate disinformation cascades. Research shows that these systems can incentivize conflict actors toward more divisive and potentially violence-inducing speech, rewarding posts that trigger anger or outrage. As a result, these types of falsehoods circulate much faster than factual corrections Vosoughi, S. et al., 2018). Platform designs that prioritise shares, comments and watch-time unintentionally boost extremist and conspiratorial material, including racist disinformation. This environment encourages the formation of echo chambers where users are mainly exposed to information that confirms their existing views (Sunstein, C.R., 1999; Pariser, E., 2011). Inside these closed loops, prejudiced narratives rarely encounter meaningful challenge and can become increasingly extreme. The outcome is a more polarised digital public sphere in which different groups inhabit separate informational realities.

Such dynamics are damaging for democratic debate. When societies split into isolated camps, reaching agreement or compromise becomes far harder. Extremist disinformation targeting minorities therefore harms not only those communities but also the quality of democratic discourse as a whole. In Europe, this dynamic has manifested in, for example, far-right online communities that circulate a continuous diet of anti-immigrant falsehoods, reinforcing their followers' xenophobic worldviews.

The political consequences can be serious. Disinformation about minorities is often weaponised during election campaigns to mobilise voters through fear and resentment. Spain's 2019 elections offer a clear example: investigators documented a surge of viral messages, many of them anti-immigrant or anti-Muslim, circulating across social media and private messaging apps (Avaaz, 2019). Some of these networks were later removed for coordinated inauthentic behaviour, but by then the damage was done – false narratives had helped shape the electoral agenda. Similar patterns have appeared since then in Spain and also across Europe: far-right and other extremist actors spread conspiratorial claims about Muslim refugees or Roma communities to gain political advantage. Elections, however, are supposed to be decided on policy choices and democratic preferences, not on fabricated stories about minority "threats".

In sum, prejudiced disinformation is not just a series of isolated hateful messages; it represents a structural attack on the foundations of democratic life. It distorts reality, splits communities, and undermines trust in the institutions that hold democratic societies together. By fueling echo chambers and deepening polarisation, it erodes the shared factual ground and mutual respect that genuine democratic debate requires. Ultimately, countering prejudiced disinformation is not only a matter of protecting minority rights – it is also essential for safeguarding the integrity and values of democratic societies as a whole.

4. Spanish case study

The Spanish information environment illustrates how prejudiced disinformation targeting minority and vulnerable groups is not merely an expression of social bias, but a strategic tool that reshapes public perceptions, fuels polarisation, and ultimately weakens democratic cohesion (Campos Domínguez, E., Esteve Del Valle, M., and Renedo Farpón, C., 2022). While Spain is not exceptional in this regard, it provides a clear example of how disinformation that appears to focus on specific communities can serve broader political and destabilising democratic objectives.

A recurring pattern in Spain involves false narratives linking immigration to crime, insecurity or welfare abuse. These narratives are often fabricated or grossly exaggerated, yet they circulate widely on social media, messaging apps, and partisan digital outlets. Their purpose is not simply to portray migrants or Muslim communities negatively, but to activate fear and resentment in the broader population. During the killing of an 11-year-old boy in Mocejón (Toledo) in 2024, for example, disinformative posts immediately claimed that the perpetrator was a Maghrebi, Muslim or Roma minor. Despite rapid confirmation by the Guardia Civil that the suspect was a Spanish national, the false claim had already reached thousands of users, generating a wave of hostility toward migrant communities and reinforcing long-standing stereotypes (Morales, E.G., August 2024). The narrative's virality, rather than its accuracy, shaped public debate in the crucial hours following the event.

This phenomenon is not limited to isolated episodes. Spanish courts have repeatedly encountered cases where disinformation is deliberately used to stigmatise minority groups. In one notable judgment, the Provincial Court of Barcelona convicted a Guardia Civil officer for spreading a manipulated video that falsely depicted an assault by an unaccompanied migrant minor. The court emphasised that disseminating such material, even without direct incitement to violence, contributed to an atmosphere of hostility and discrimination that merited criminal sanction (nº674/2022, November 8th). These cases demonstrate the capacity of prejudiced disinformation to turn marginal actors into influential vectors of hatred, often supported by algorithmic amplification.

The effects of these narratives are visible in opinion data. According to the September 2024 CIS Barometer, immigration suddenly became one of Spain's top three perceived national problems for 30% of respondents, up from 11% only three months earlier. While fluctuations in survey responses can have multiple causes, the sharp rise coincided with a period in which monitoring bodies detected a significant increase in potentially disinformative content targeting migrants and racialised communities. The temporal alignment suggests that disinformation does not merely reinforce existing prejudices but can actively reshape public priorities and fuel political anxieties.

Other minority groups have also been targeted. LGBTQ+ communities have been subject to recurring waves of disinformation (Shevtsova, M., 2020), particularly around symbolic dates such as Pride or during debates surrounding the 2023 "Trans Act". Narratives portraying LGBTQ+ people as threats to children or as agents of "gender ideology" have circulated across messaging apps and fringe media channels. Some of these narratives originate domestically, while others mirror messaging identified by the EU's External Action Service (2023) as part of broader foreign influence operations. Even when the immediate target appears to be a minority group, the strategic aim is broader: to undermine trust in institutions, weaken support for equality measures, and reinforce divisive identity politics.

The Spanish experience also shows how prejudiced disinformation escalates during moments of political tension. Electoral cycles, in particular, tend to coincide with spikes in misleading narratives centred on migration, Islam, and gender-related issues. These tactics leverage emotional triggers – fear, disgust, or moral outrage – to mobilise voters or delegitimise opponents. Their purpose is not simply to attack the targeted group but to reorient public debate toward cultural conflict and away from structural socio-economic issues.

Taken together, these examples reveal that prejudiced disinformation in Spain functions as more than hate-driven content: it is a mechanism for destabilising democratic discourse. By exploiting pre-existing biases, it fractures social cohesion and diminishes trust in public institutions. Its strategic use by political actors and, at times, foreign influence networks shows that the ultimate objective often extends beyond harming a specific minority. Instead, the aim is to reshape the political agenda, deepen polarisation, and weaken the epistemic foundations that democratic decision-making depends on. Prejudiced disinformation is therefore both a direct threat to minority dignity and a structural threat to democratic resilience.

5. Regulatory framework and responses

The European Court of Human Rights has drawn a firm line against extreme hate speech targeting minorities, even where there is no direct incitement to violence. In Norwood v. United Kingdom (2004), a man displayed a poster vilifying all Muslims ("Islam out of Britain" with an image of the burning Twin Towers). The Court deemed this extreme anti-Islam expression an abuse of rights incompatible with the Convention's values of tolerance and social peace, and thus not protected by freedom of expression. In Vejdeland and Others v. Sweden (2012), the Court upheld the convictions for distributing homophobic leaflets in a school, although the leaflets did not explicitly call for violence. The judges stressed that inciting hatred does not require a call for violence – insulting or slandering a group can be enough for authorities to sanction a speaker in order to protect the rights of others. They also noted that discrimination based on sexual orientation is as serious as that based on race or religion.

Similarly, in Féret v. Belgium (2009), involving a politician's xenophobic anti-immigrant pamphlets, the Court held that his hate-speech conviction did not breach Article 10. Such rhetoric from an elected official was seen as a danger to social peace and stability in a democratic society, warranting criminal sanctions. The Court underscored that the absence of a direct call to violence does not excuse hate speech: even without explicit incitement, states can legitimately restrict expressions that spread, incite or justify hatred. It further emphasized that politicians, especially, must avoid fomenting intolerance – their public pronouncements should not promote racial or ethnic hatred, given their influence and the threat such speech poses to social cohesion.

EU Legal Framework: Hate Speech and Disinformation

At the European Union level, the approach distinguishes between illegal hate speech and harmful but lawful disinformation. The key instrument against hate speech is Framework Decision 2008/913/JHA, which obliges Member States to criminalize "public incitement to violence or hatred" on grounds such as race, color, religion, descent, or national/ethnic origin (article 1), thereby establishing a common baseline outlawing racist and xenophobic expression across the EU.

In 2016, the European Commission and major IT companies introduced an EU Code of Conduct on Countering Illegal Hate Speech Online, under which platforms pledge to remove illegal hate speech within 24 hours. This voluntary code, monitored in cooperation with NGOs, has markedly increased the takedown rate for unlawful hate content on social media.

For online disinformation, which generally is not illegal, the EU has relied on voluntary cooperation. In 2018 the Commission launched a Code of Practice on Disinformation (strengthened in 2022), through which leading online platforms and advertisers commit to curb the spread of false or misleading content. Signatories of this code implement measures such as demonetizing disinformation (cutting advertising revenues to purveyors of false news) and disabling fake accounts and bots that amplify false narratives. The European Commission credits this self-regulatory approach with helping contain recent disinformation surges.

In 2022, the EU enacted the Digital Services Act (DSA), a binding regulation that reinforces these initiatives and draws a clear line between unlawful and lawful content. The DSA requires online intermediaries to expeditiously remove illegal content, including hate speech outlawed under national or EU law, once notified. At the same time, the DSA addresses "harmful but legal" content through transparency and accountability rules rather than direct censorship. Very large online platforms must assess the systemic risks of phenomena like viral disinformation on their services and "take reasonable measures" to mitigate those risks. In practice, this means they must examine how their algorithms may be amplifying harmful falsehoods or extremist content and adjust their systems to reduce those impacts. The DSA thus complements the EU's hate speech laws by requiring oversight and risk management for online mis and disinformation, aiming to strike a balance between combating digital harms and respecting lawful expression.

Spanish Legal Framework

Spain's domestic legal framework implements these principles with strict criminal provisions against hate speech and additional civil/administrative measures. The centerpiece is Article 510 of the Criminal Code, which criminalizes public incitement to hatred, hostility, discrimination or violence against protected groups defined by characteristics such as race, religion, ethnicity, national origin, gender, or sexual orientation. Article 510 also prohibits related acts: it punishes disseminating written or other material that incites hatred, the denial or gross trivialization of genocides or crimes against humanity when that creates an atmosphere of hatred, and serious insults or slurs that demean people due to their group identity. Spanish courts have clarified that the hate speech law targets only grave attacks on vulnerable communities, ensuring that it protects minority groups without unduly limiting legitimate discourse or criticism that lacks a hateful intent (Spanish Constitutional Court, 235/2007; Spanish Supreme Court, 123/2017; Spanish Supreme Court, 72/2018; Spanish Supreme Court, 489/2020).

Beyond the criminal law, Spain uses civil and administrative tools to combat hate speech and uphold equality. Anti-discrimination statutes allow victims to seek relief (for example, damages or injunctions for collective hate incidents), and authorities can impose administrative sanctions for hate-related offenses (for instance, racist behavior at a sports event may result in fines or bans). Public agencies also emphasize prevention and monitoring: the Interior Ministry publishes annual reports on hate crime trends, and official protocols and training guide police and prosecutors in identifying hate speech.

Spain has likewise moved to counter disinformation. In October 2020, the government issued Order PCM/1030/2020, which established a national Procedure for Action Against Disinformation. This order created coordination mechanisms to detect and respond to disinformation campaigns (especially foreign interference) in line with the EU's Action Plan on Disinformation. It enables Spanish institutions to participate in EU alert systems and to mount timely responses when online falsehoods threaten public order or democracy. Spain's approach treats disinformation as a security challenge managed through interagency cooperation and public awareness, rather than through criminal prohibition of content.

6. Rethinking responses: from content moderation to democratic resilience

Given the systemic nature of prejudiced disinformation, responses limited to punitive legal measures or platform-level content takedowns are insufficient. What is needed is a shift from treating disinformation as a problem of "false content" to understanding it as a structural threat to democratic resilience, social cohesion, and minority protection.

First, states must recognize that prejudiced disinformation rarely exists in isolation. It intersects with electoral manipulation, foreign influence operations, and long-standing discriminatory narratives. Effective responses must therefore address the broader ecosystem: the political incentives that encourage inflammatory speech, the economic models of social media platforms that privilege engagement over accuracy, and the historical biases that make certain groups especially vulnerable to manipulation.

Second, regulatory frameworks should be complemented by strong preventive and educational measures. Media literacy programmes, public awareness campaigns, and transparent communication from institutions are essential for reducing the susceptibility of citizens to disinformation. Empowering users with critical digital skills is particularly important in societies where algorithmic amplification can turn fringe narratives into mainstream controversies within hours.

Third, platform accountability must go beyond voluntary commitments. The Digital Services Act represents an important step in requiring large platforms to assess systemic risks and adjust algorithmic design accordingly. However, meaningful enforcement—accompanied by independent audits, transparency requirements, and sanctions for non-compliance—is crucial to ensure that platforms reduce the virality of harmful content rather than merely removing illegal material after the fact.

Fourth, protecting minorities requires robust legal safeguards and proactive monitoring. Hate-speech laws, such as Spain's Article 510, play a significant role, but legal interventions should be paired with support systems for targeted communities, including civil-society organisations, anti-racist watchdogs, and community-based reporting mechanisms. These actors provide early detection of narrative shifts and help mitigate the psychosocial harm that disinformation campaigns inflict.

Finally, any long-term strategy must explicitly acknowledge that prejudiced disinformation can serve geopolitical aims. Foreign influence operations frequently exploit societal divisions, using minority-targeted falsehoods to undermine democratic trust and disrupt political stability. European democracies must therefore incorporate disinformation into their national security frameworks, as Spain has done through its Procedure for Action Against Disinformation (Order PCM/1030/2020). Strengthening cooperation between national authorities, EU institutions, and civil society is essential to building a coordinated and multilayered response.

In conclusion, countering prejudiced disinformation requires moving beyond reactive content policing toward a holistic model of democratic resilience. Such an approach recognizes that disinformation harms not only the dignity and safety of minority communities but the very foundations of democratic life: informed citizenship, pluralism, and institutional trust. Protecting vulnerable groups and safeguarding democracy are therefore inseparable tasks: responding to prejudiced disinformation is not merely a matter of minority rights, but a core democratic imperative.

References

Aba Catoira, A. M. (2020). Los desórdenes informativos en un sistema de comunicación democrático. *Revista de Derecho Político*, 1(119), pp. 119-151. DOI: https://doi.org/10.5944/rdp.109.2020.29056

Arcila-Calderón, C., Blanco Herrero, D., y Valdez Apolo, M. B. (2020). Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español. *Revista Española de Investigaciones Sociológicas*, 172, pp. 21-40. DOI: http://dx.doi.org/10.5477/cis/reis.172.21

Arcila Calderón, C., Sánchez Holgado, P., Gómez, J. et al. (2024). From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanit Soc Sci Commun* 11, pp. 1-14. DOI: https://doi.org/10.1057/s41599-024-03899-1

Arias Maldonado. M. (2016). *La democracia sentimental: política y emociones en el s. XXI*. Barcelona: Página indómita.

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Campos Domínguez, E., Esteve Del Valle, M., and Renedo Farpón, C. (2022). Retóricas de desinformación parlamentaria en Twitter. *Comunicar*, (72), pp. 47-

58. DOI: https://doi.org/10.3916/C72-2022-04

Chadwick, A. (2017). *The hybrid media system: Politics and power* (2nd ed.). Oxford University Press.

Dahl, R. (1956). *A Preface to Democratic Theory*. Chicago-London: The University of Chicago Press.

Foucault, M. (1991). Governmentality. In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect: Studies in governmentality*, (pp. 87–104). University of Chicago Press.

Freund, J. (1968). *La esencia de lo político* (S. Nöel, translation). Madrid: Editora Nacional.

Fuentes Lara, C. and Arcila Calderón, C. (2023). Islamophobic hate speech on social networks: An analysis of attitudes to Islamophobia on Twitter. *Revista Mediterránea de Comunicación*, 14(1), pp. 45–58. DOI: https://doi.org/10.14198/MEDCOM.23044

Gilroy, P. (2004). *After empire: Melancholia or convivial culture?* Routledge.

Montalto Monella, L., and Palfi, R. (2020, March 3). Orban uses coronavirus as excuse to suspend asylum rights in Hungary. *Euronews*. Available online at: https://www.euronews.com/2020/03/03/orban-uses-coronavirus-as-excuse-to-suspend-asylum-rights-in-hungary

Morales, E. G. (August, 2024). Asesinato en Mocejón: Así operan los ultras para expandir odio con bulos racistas en redes. *Público*. Available online at: https://www.publico.es/sociedad/asesinato-mocejon-asi-operan-ultras-expandir-odio-bulos-racistas-redes.html

Office of the United Nations High Commissioner for Human Rights. (n.d.). A*bout minorities and human rights*. Available online at: https://www.ohchr.org/en/special-procedures/sr-minority-issues/about-minorities-and-human-rights

Office of the United Nations High Commissioner for Human Rights. (2020, March 9). *Visit to Spain: Report of the Special Rapporteur on minority issues* (A/HRC/43/47/Add.1). Available online at: https://www.ohchr.org/en/documents/country-reports/ahrc4347add1-visit-spain-report-special-rapporteur-minority-issues

Servicio Europeo de Acción Exterior (SEAE). (2023). *Informe sobre amenazas de manipulación y desinformación de información extranjera (FIMI) dirigidas a las comunidades LGBTIQ+*. División de Stratcom del SEAE. Available online at: https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-LGBTQ-Report-03-Digital%201.pdf

Shevtsova, M. (2020). Fighting "Gayropa": Europeanization and Instrumentalization of LGBTI Rights in Ukrainian Public Debate. *Problems of Post-Communism,* 67(6), pp. 500–510. DOI: https://doi.org/10.1080/10758216.2020.1716807

Strand, C. and Svensson, J. (2021). *Disinformation campaigns about LGBTI+ people in the EU and foreign influence*. European Parliament, Policy Department for External Relations, Directorate-General for External Policies. Available online at: https://dspace.ceid.org.tr/xmlui/handle/1/1805

Sunstein, C. R. (1999). The law of group polarization. *John M. Olin Program in Law & Economics Working Paper*, 91.

Szakács, J. and Bognár, É. (2021). *The impact of disinformation campaings about migrants and minority groups in the EU*. European Parliament, Directorate-General for External Policies.

Taylor, A. (2016, January 29). An alleged rape sparked tensions between Russia and Germany. Now police say it was fabricated. *The Washington Post*. Available online at: https://www.washingtonpost.com/news/worldviews/wp/2016/01/29/an-alleged-rape-sparked-tensions-between-russia-and-germany-now-police-say-it-was-fabricated/

The Poynter Institute. (n.d.). *IFCN Covid-19 misinformation*. Available online at: https://www.poynter.org/ifcn-covid-19-misinformation/

Vosoughi, S., Roy, D., y Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380).

Wardle, C. y Derakhshan, H., (2017) *Council of Europe Report [DGI (2017)09] on Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe publications.

World Economic Forum. (2017). *Global Risk Report (Report nº 12)*. Available online at: http://wef.ch/risks2017

Fact-checking as a tool to tackle hate speech: A systematic literature review

Juliana da Cunha Mota, University of Oxford

Introduction

Social media platforms were originally perceived as a tool for democratising access to knowledge and information. Especially from the 2010s, however, they have become a particularly fertile ground for propagation of mis- and disinformation and hate speech (Matamoros-Fernández & Farkas, 2021; Schradie, 2019).

Hate speech and misinformation are two distinct phenomena that cause different harms (Heldt, 2019). Yet, they are often connected: during crises, misinformation can support, justify, or fuel hateful content – creating what we call 'populist hate narratives' (Erjavec & Kovačič, 2012; Mayagoitia-Soria et al., 2024; Poole et al., 2021).

Various tools and initiatives are employed to tackle hate speech and misinformation. AI and machine learning are increasingly used in the fight against hate speech (Jahan & Oussalah, 2023). To tackle mis- and disinformation, fact-checkers (FCs) play a crucial role (Adam, 2025; Graves, 2017; Westlund et al., 2024). Given FCs' importance in reducing online disinformation, one could reasonably assume that they should also contribute to reducing the spread of hateful content. Nonetheless, the role of FCs in tackling hate speech through fact-checks of disinformation remains underexplored by the literature.

This paper seeks to assess the existing evidence on the role of FCs in tackling the spread of hate speech on platforms. Through a systematic literature review and critique of academic articles on hate speech and fact-checking published within the past decade, the paper aims to answer the following research question: What does the literature (2015-2025) report about how collaborations between fact-checkers and platforms are associated with the prevalence of populist hate narratives?

To answer this question, I performed a qualitative analysis of published material. I reviewed the literature to assess the existing evidence and scholars' perceptions of the following issues: (1) the relationship between hate speech and disinformation, (2) the role of FCs in tackling hate speech; (3) the role of platforms in facilitating or curbing the dissemination of hateful narratives.

By delving into the role of FCs for tackling hate speech, this paper makes three contributions. First, by identifying the gaps in the scholarship, it assesses areas that can benefit from further research. Second, by expanding the results of this study, it considers the consequences of scaling down on collaborations between platforms and FCs. This is a particularly timely contribution in light of Meta's decision to end the 3PFC in the US (Kaplan, 2025). While questions remain if Meta will replicate the measure elsewhere, the results of this analysis can provide useful insights into the impact of this measure. Finally, the results will guide broader legal discussions, contributing to the growing debate on States' obligations concerning electoral disinformation and/or positive obligations to tackle hate speech (Alkiviadou, 2025; Pentney & Shattock, 2025).

This paper proceeds as follows: Section 2 starts by expanding on the existing literature on the role of FCs and FC-platforms collaborations. Section 3 presents the method employed (systematic literature review) and its justification, further explaining the data collection. Section 4 contains my main findings. Finally, Section 5 expands the findings to propose broader discussions guided by the discussion question.

Fact-checkers and collaborations between FCs and platforms

FC emerged as a global, yet fragmented, movement. FC firms emerged in the 1990s to hold politicians and other public figures accountable for false statements. Originally incubated in US newsrooms, the movement grew and changed as social media platforms challenged journalists' gate-keeping role and mis- and disinformation became more prevalent. As the literature indicates, 'In response to disinformation campaigns during the 2016 US presidential election, the field's focus shifted from verifying claims made by politicians to policing viral misinformation on digital platforms – the debunking turn'(Mahl et al., 2024, p. 3).

With the debunking turn, FC is now performed by different actors, including but not limited to professional journalists, political campaigns and party organisations, and third sector organisations. Each of these actors have different perceptions about the role of FCs, performing it by different standards and methods (Cavaliere, 2020, p. 158). Despite these differences, fact-checkers have somewhat similar goals: to verify information presented as a fact. They do not verify statements of opinion or statements which cannot be corroborated. Other similarities involve their areas of practice, which usually comprise: choosing claims to check, contacting the speaker, tracing false claims, dealing with experts, and showing your work (Graves, 2017). FCs' work supports both individual readers and content moderation on large-scale platforms through labelling mechanisms (Sehat et al., 2024).

Acknowledging the importance of FCs, platforms started collaborating with FC organisations in the last decade. For instance, Meta instituted its third-party fact-checking programme ('3PFC') in 2016 'to reduce the spread of misinformation and provide more reliable information to users' (Bengtsson et al., 2025, p. 249). Similarly, Google has partnered with FCs to develop data standards to surface fact-checks in search results. Some FC organisations choose to join these partnerships to amplify the reach of their content (Bélair-Gagnon et al., 2023).

The importance of FC-platforms collaborations cannot be overstated. Some FC organisations exist and operate thanks to platform partnerships. However, in January 2025, Meta announced it was scaling down on these partnerships in the US. The measure was taken under the guise of strengthening the protection to freedom of expression, allegedly threatened by FCs who were, in Mark Zuckerberg's words, 'too politically biased'(Kaplan, 2025). In FCs' view, the information ecosystem of several countries could be affected should Meta decide to expand its decision to scale back FC collaborations globally. This is because Facebook and Instagram are the main sources of news in countries from the global south. Thus, the lack of fact-checking initiatives could increase the circulation of misleading information (Kahn, 2025).

Scholars in media studies have explored how FCs and platforms conceive making trade-offs in their platforms to build their fields as agents of knowledge. They found that platforms and FCs are marked by both asymmetric and mutual dependence. They also recognise the need to acknowledge the inter-relational and dynamic element of platform companies, and how they relate to a larger platform and information ecosystem (Bélair-Gagnon et al., 2023). While explored in other areas of knowledge, the relationship between FCs and platforms remains under-explored by legal and socio-legal scholars.

Method

To explore FC-platforms collaborations, I employed a systematic literature review of the existing works on the topic. The aims of this work justify choosing standalone systematic literature reviews (SSLR) as a method. Specifically, this paper aims at (1) answering the research question based on the existing scholarly evidence; and (2) identifying gaps in the literature. These are precisely the same aims as those of SSLR (Okoli, 2015; Petticrew & Roberts, 2008). Furthermore, unlike scoping reviews, systematic SSLR are somewhat targeted. For instance, they may aim at identifying the evidence on the efficacy of a specific intervention, instead of broadly exploring the characteristics of an area of knowledge (López-Borrull & Lopezosa, 2025). Thus, SSLR is in line with the aims of this paper.

Importantly, SSLR is not valuable in the early days of a research field, which further justifies its usage in this specific research. In this paper, I assess a decade of the body of literature on misinformation, fact-checking, and hate speech. The choice for setting 2015 as the cutoff date for this analysis is twofold. First, the Trump election and Brexit happened in 2016 and 2017, respectively. These two episodes triggered debates into the role of platforms in spreading mis- and disinformation, especially during electoral periods. Second, Meta established the 3PFC in 2016 and the programme grew bigger in the following years. Accordingly, analysing ten years of corpus literature into misinformation, hate speech, and fact-checking from 2015 to 2025 might reveal how scholars perceived these topics in light of societal and technological developments, if at all.

SSLR require a systematic methodological approach, explicit explanation of the procedures followed, and a comprehensive scope of the material (Okoli, 2015). Below, I provide an overview of the methodical approach adopted and the procedures followed.

1. Data collection

The first step into the data collection phase consists of choosing the appropriate database of academic works. In line with existing literature, I chose to extract works from Scopus, given its prestige and quality of papers (López-Borrull & Lopezosa, 2025). Admittedly, choosing one database limits the results of this paper. Nonetheless, due to time and scope constraints, performing an in-depth analysis of other databases was not feasible.

After selecting the database, I defined the search terms. I performed several searches with the following keywords:

Search	Keywords
1	'hate speech' and 'fact-checking'
2	'hate speech' and 'fact-checkers'
3	'hate speech' and 'fact-checkers' and 'social media platforms'
4	'hate speech' and 'fact-checkers' and 'platforms'
5	'hate' and 'fact-checking'
6	'hate' and 'fact-checkers'
7	'hate' and 'fact checkers' and 'platforms'
8	'hate speech' 'misinformation' and 'elections'
9	'hate speech' and 'fake news' and 'elections'
10	'hate speech' and 'fake news' and 'platforms'
11	'hate speech' and 'populism' and 'platforms'
12	'hate speech' and 'populism'
13	'hate speech' and 'misinformation'

I considered solely papers published from 2015 to 2025, for the reasons mentioned in the section above. As illustrated by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram shown in Figure 1, our searches returned n=3,375 works in total (Gibson et al., 2025; Liberati et al., 2009).

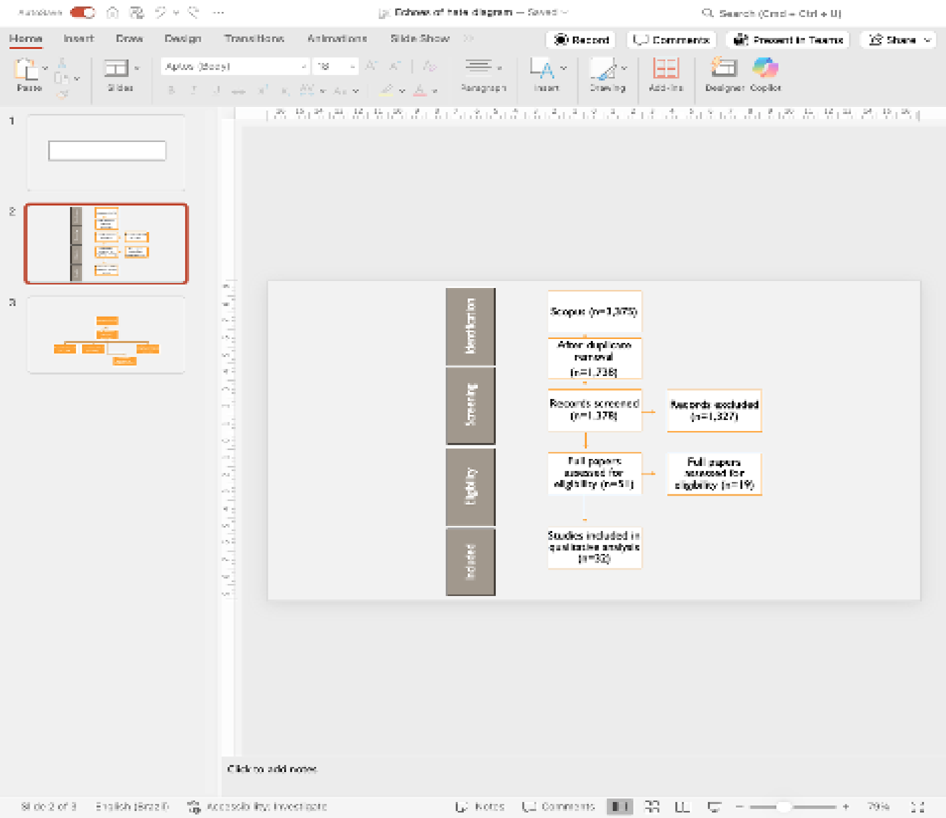
I uploaded the full database of papers into Ryann.AI, a software for systematic literature review. I removed 1,637 papers which were duplicated, after which I conducted the first round of revision. In this round, I considered solely the title of the papers, the abstract, and the key words. I excluded (i) books and chapters; (ii) conference proceedings. Furthermore, I also did not consider papers that (i) focus on countries outside of Europe; (ii) in languages other than English; (iii) which had not been peer-reviewed; and (iv) focused on the role of traditional media or dark web forums in disseminating hate speech, without further consideration to social media platforms. This round resulted in the exclusion of 1,686 works. Subsequently, I performed a second round of revision, mainly to exclude book chapters and conference proceedings. After the second round of exclusion, I had a preliminary database of works constituted of 51 papers.

All these 51 papers were downloaded and assessed. Some papers were subsequently excluded as for similar reasons as those previously exposed (i.e. they focused on the role of traditional media, delved into content moderation issues through removal of content and not fact-checking, or concerned models for fact-checking, mis- or disinformation, without exploring hate speech issues or the role of platforms). Admittedly, the exclusion of grey literature may have led to the omission of potentially relevant works. However, this limitation does not compromise the validity of the study, as the author believes theoretical saturation was reached with the papers included. Finally, the database of examined papers included 32 works. A full list of papers considered is provided in the Appendix A.

Figure 1: PRISMA flowchart

Findings

The analysis of the papers revealed insights about the type of analysis being conducted in the fields of hate speech, misinformation, FCs, and platforms. Below, I present some of these findings.



Identification

Screening

Eligibility

Included

Scopus (n=3,375)

After duplicate removal (n=1,738)

Records screened (n=1,378)

Records excluded (n=1,327)

Full papers assessed for eligibility (n=51)

Full papers removed for eligibility (n=19)

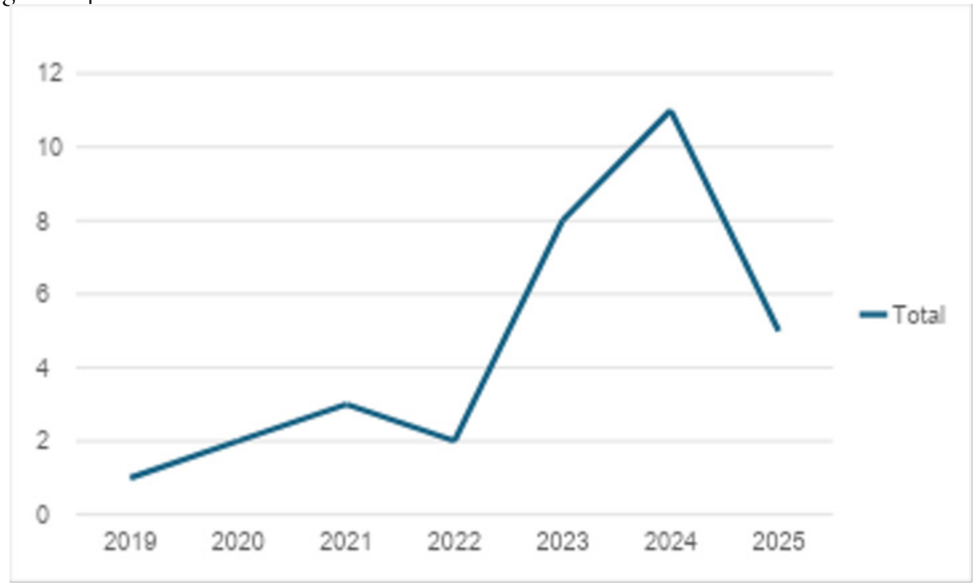
Studies included in qualitative analysis (n=32)

Click to add notes

1. Growing interest in the topic

There appear to be growing interest for the topics of hate speech, misinformation, platforms, and FCs. No eligible studies appear prior to 2019, with outputs rising and peaking in 2024.

Some political and social factors might explain the surge in papers from 2022 onwards. First, as previously mentioned, platforms' actions and omissions to tackle misinformation came to the centre stage from 2016 onwards, with the Trump election, the Brexit Referendum, and the Cambridge Analytica scandal. Subsequently, the COVID-19 further highlighted the role of platforms in tackling hate speech and misinformation.



2. Empirical vs theoretical analyses

Most of the papers consisted of empirical analyses of the issues, often adopting quantitative and qualitative methods to assess harmful content. Theoretical works are scarce, fewer than 10 papers. Yet, these theoretical contributions provided invaluable insights which should be accounted for.

Most of the empirical papers focused on one specific type of hate speech and platform for analysis. For instance, Asardag (2025) employed critical feminist lens to explore the discourses around Meta's changes to their hate speech policies, announced in January 2025. She found that the changes would likely negatively impact LGBTQAI+ people, consistently targeted by online hate speech and disinformation.

Other papers empirically analysed disinformation and hate speech. These studies are crucial for understanding how hate speech spreads within different contexts and platforms. Conversely, as Liu et al. had previously noticed, most of these studies are limited to specific cases, and there is a lack of longitudinal and generalised study of hate speech (Liu et al., 2024). Understandably, hate speech is context-dependent, thus it is questionable whether an overarching empirical study would produce significant and adequate results. For instance, Carrasco-Farré (2022) conducted a computational linguistics assessment of misinformation and harmful content, Cinelli et al. (2021) considered hate comments on YouTube, and Dowining & Dron (2020) focused on discourse analyses of tweets.

While in the minority, theoretical papers were also important for drawing connections between disinformation and hate speech. For instance, Cepollaro et al.'s work theoretically assessed efficacy and deontic questions of counterspeech for tackling harmful messages. Ultimately, they argued that philosophers should focus more on the efficacy question to advocate for better frameworks to conceptualise the efficacy of counter speech. A paper linking these philosophical views and empirical works was, however, missing from our database of papers (Cepollaro et al., 2023).

3.Platforms investigated

Most of the empirical papers focused solely on some platforms for analysis. The most popular platform was X (formerly Twitter), which accounted for 9 papers, followed by Facebook (with 7 papers), TikTok (3 papers), and YouTube (2 papers).

Twitter's API for research purposes might explain the platform's popularity among researchers. Until 2023, the platform was 'a playground for academic research', providing access to millions of tweets every day for free. After Twitter's acquisition by Elon Musk, the policies changed and access to the API is now behind a paywall (Sarah Grevy Gotfredsen, 2023). While there were four papers published in 2024 using Twitter's data, I found only 2 in 2025. This raises questions whether we will continue to see a decrease in works with Twitter's data because of Musk's new policies.

Another interesting point concerns papers analysing TikTok content. Scholarly attention to the platform did not appear to accompany its surge in users and importance (Pérez Rastrilla et al., 2023). It is unclear why scholars paid so little attention to TikTok and whether this platform will appear more prominently in future studies.

4.Exploring links between hate speech and disinformation

Not all hateful messages contain disinformation, and vice versa. Yet, numerous studies recognised the connections between disinformation, hate speech, radicalisation, and polarisation.

For instance, according to Carrasco-Farré's (2022) computational linguistics study, false information and hate speech share some commonalities, including the negativity of the content and their elevated appeal to moral values. Similarly, Mohsen et al. (2024) applied linear regression models with standardised coefficients to assess a database of more than 8 million tweets from N=6,832 Twitter users, revealing a link between misinformation and harmful language. Echoing this, Zhou et al. (2025) found that 28% of the hate speech instances against Asians contained misinformation.

Conversely, some studies found no relation between hate language and misinformation communities. For instance, Cinelli et al. (2021) did not find any correlation between the usage of hate language by YouTube users and their involvement in misinformation communities in the platform. Downing and & Dron (2020) echoed the finding that social media was not a conduit for hateful or deceitful messages. Their analysis focused on Tweets about the Grenfell fire and the Muslim community, and they found that the most influential Tweets pictured Muslims in a positive light. Poole et al. had similar findings when analysing tweets containing the hashtag #StopIslam. In their study, the most shared tweets contained counter-narrative and positive messages about Islam instead of hateful content(Poole et al., 2021).

In summary, with some exceptions, I found evidence in the existing literature that there are links between misinformation and hate speech online – the extent of these links is, however, disputed.

5.Exploring the role of FCs in tackling hate speech

I found few papers acknowledging the importance of FCs or fact-checking techniques in tackling misinformation and hateful content online. For instance, Miškolci et al. (2020) adopted a quasi-experimental research design to dissect the spread of anti-Roma discourses on Facebook. In the study, the authors adopted the role of FCs to assess the consequences of entering pro-Roma comments, often checking facts presented in posts targeting this minority. They found a positive correlation between their positive comments and new users' interventions. Put differently, when they made posts with pro-Roma comments, it motivated 'other followers of those particular Facebook profiles to join the discussion arguing in favour of the Roma as well'(Miškolci et al., 2020, p. 139). Conversely, they acknowledge that fact-checking as a counterspeech strategy showed limitations, as most of the users continued to post hateful comments within the thread, while others rarely acknowledged the argument from their intervention. Finally, the authors highlighted that emotions spread through Facebook might have also contributed to users' comments and reactions. As such, the authors concluded that platforms' structural design may contribute to creating communities of hate and a revision of this business model is needed to tackle hate speech (Miškolci et al., 2020).

While some papers acknowledged the importance of fact-checking in tackling hate narratives, Munn (2024) highlighted some limitations to this approach. According to him, the existing approaches to tackle misinformation (including FCs) are based an idealised rational version of humans, which fails to consider that humans are primarily, rational, factional, and bigoted. Accordingly, they argued that more information will not solve the misinformation issue. Instead, the author proposed linking fact-checking approaches to psychological research to robustly connect human insights into real-world interventions.

6.Exploring the role of platforms in disseminating or curbing hate narratives

Finally, some papers addressed the role of platforms in curbing or disseminating disinformation and populist hate narratives. For instance, Munn (2024) highlighted that humans are emotional beings, often acting driven by personal and/or moral values rather than by objective evidence. Online platforms, in turn, capture and amplify this emotion through algorithms that privilege controversial or attention-seeking content – including populist and hate narratives.

Silva & Parker (2025) argued that platforms' structural power in speech dissemination impedes their regulation. According to them, platforms use discursive power to maintain control over the conditions for their own regulation. In other words, platforms discursively claim that they constitute the ones, or most legitimate ones, who should serve as the arbiters of the speech they host. Consequently, they are central in disseminating or curbing hate speech. Heldt (2019) echoes this.

Echoing the argument that platforms play an active role in choosing which content to disseminate, Abdul Reda and Alkhonin (2025) found that 'algorithms could be adjusted (…). This strategic shift could reduce the space available for misinformation to spread, aligning public focus with verified informative content continuously' (Abdul Reda & Alkhonin, 2025, p. 18). Similarly, Silva & Parker (2025) claimed that platform power can directly contribute to the proliferation of hate speech when they fail to remove content. In doing so, they proposed a typology, building on Fuchs (2013) for platforms' actions in contributing to hate speech: amplification, accommodation, or discursive authorisation of hate speech.

The analysis of platforms' terms and services further highlighted their active role in curbing hate narratives (Arora et al., 2024). Despite this, several studies found hateful content in the platforms analysed. For instance, González-Aguilar et al. (2023) found few hateful messages on TikTok in a study assessing populist speech on the platform. The authors highlighted that hate speech played no particular influence on the videos' views (González-Aguilar et al., 2023). While somewhat surprising, the authors acknowledged one important caveat to their study: TikTok might have played an important role in downgrading videos with hateful content, thus affecting the videos' views. Therefore, the platform might have become a tool and strategy in downplaying harmful content.

Poole et al. (2021) echoed the argument that platforms play a crucial role in disseminating or curbing populist hate narratives. They argued that Twitter played an active agent role in setting the narrative surrounding the use of the hashtag #StopIslam. This is because Twitter (now X) had removed a significant number of tweets containing this contested hashtag. The authors also found that Twitter reinforced the dominance of the 'elite', by using algorithms that amplified the visibility and reach of content posted by influencers. Thus, they claimed that social media platforms' dynamics afforded more agency to well-organised groups with stronger ties. In their study, these groups were aligned with the right, fuelled by populist discourse. Accordingly, they concluded that 'What is particularly concerning about #stopIslam, therefore, is that it illustrates how the strategies of social media platforms can create conditions that lend themselves not just to the actions, but ideological commitments of right-wing populist groups' (Poole et al., 2021, p. 1438).

While I did not find studies generally disputing platforms' role in spreading hate speech, I found at least one study challenging pre-established conceptions about the dissemination of harmful speech. Among others, Budak et al. (2024) challenged three media claims. First, that disinformation is growing in the age of social media and that more people are exposed to it. Second, that exposure to misinformation and harmful content is primarily driven by the algorithms of platforms. Finally, that there are correlations between exposure to harmful content online and undesirable psychological or behavioural effects. According to them, the existing studies on online harmful speech do not support any of these conclusions. In fact, they argue that the literature demonstrates that the exposure to misinformation is low, often concentrated among a small minority. Similarly, they contend that users are not fed harmful content through 'filter bubbles' – instead, users who are already attentive to this kind of content harmful content seek it out across mediums.

In light of the above, the existing evidence on the scholarly literature points to a link between the platforms and the dissemination of hateful messages. If platforms' architecture contributes to the dissemination of populist discourses and hate speech, the question is how FCs impact this ecosystem, if at all. Sehat et al. (2024)'s study found that platforms indicated which content they wanted prioritised for fact-checking. (Sehat et al., 2024). Thus, platforms could arguably indicate the need to have populist hate narratives checked.

In sum, it is evident that platforms actively try to curtail hate speech, which often serves as a basis for populist narratives. However, their approach is arguably different when it comes to mis- and disinformation, with platforms historically taking a stance in favour of a free marketplace of ideas. Analysing the T&Cs of 42 platforms, Arora et al. (2024) found that at least 10 of them did not address misinformation. Furthermore, the sum of explicit mentions to misinformation in all T&Cs analysed taken together was significantly narrower (N=231) in comparison to hate speech (N=513) or violent content (N=620). As the authors themselves acknowledge, the number of mentions to a particular issue in a T&C does not perfectly encapsulate how important a topic is to a platform. However, it is a useful proxy to demonstrate how much attention platforms pay to a particular topic and the importance of detecting infringing content.

Discussion

In short, the literature showed both theoretical and empirical evidence that platforms play a central role in enabling or restraining populist hate narratives through their algorithms and terms and conditions. Regrettably, I found no papers examining FC-platforms collaborations, especially on the effects of such collaborations in curbing or enabling harmful content. This is a notable gap in the literature, which should readily be addressed.

Considering this gap, I found no compelling and clear evidence in the literature on how FC-collaborations are associated with the prevalence of populist hate narratives. Notwithstanding this, having broadly explored the existing evidence in the literature on the links between misinformation, hate speech, the role of FCs and that of platforms, I make a second attempt at answering this question by extrapolating the arguments and evidence.

1.Do platforms restrain populist hate narratives through collaborations with FCs?

As mentioned, I did not find any works focusing on the relationships between FCs and platforms, and the impact of these collaborations to the broader hate speech context. This absence of direct evidence necessitates an inferential approach, drawing from adjacent findings on misinformation and hate speech dynamics.

First, the evidence points to platforms' economic incentives to develop algorithms which prioritise divisive content, as this type of content tends to catch users' attention. The types of divisive content available online is not, however, unfettered. Notably, the prevalence of hate speech in a given platform can lead to user avoidance. Consequently, the largest platforms in the market explicitly prohibit the dissemination of hate speech through their terms and conditions.

At first, therefore, the evidence points to a reduced importance of collaborations between FCs and platforms; if platforms are strict in their policies when it comes to hate speech, then FC collaborations play virtually no part in enabling or restraining hate narratives - because the content will be removed anyway. Conversely, if the content is more nuanced and does not constitute prima facie hate speech, FCs can play a crucial role in providing further context.

Further evidence also contributes to the argument that FCs and platforms should collaborate to reduce the spread of hate speech online. First, even if platforms enforce their terms and conditions and filter hateful content, Budak et al. (2024) demonstrated they still show harmful content to users who seek it out. Drawing from this, fact-checking initiatives should contribute to play correcting false or misleading statements and avoiding further radicalisation of these users actively looking for harmful content.

Conversely, not all scholars are optimistic about the corrective potential of fact-checking. According to Munn (2024), collaborations between platforms and FCs would not necessarily restrain the spread of hateful content, as the theoretical target of these interventions is a rational human being as opposed to the actual target of these interventions, which are irrational humans. Notwithstanding, he acknowledged that FCs could contribute to reducing misinformation, if linked to psychological research on how to present facts to irrational beings. As such, he admitted that connecting human insights into real-world interventions potentially contributes to reducing misinformation.

Extrapolating the existing evidence, collaborations between FCs and online platforms arguably have the potential to mitigate populist hate narratives. Although, the degree of efficacy of such collaborations might depend on the engagement with users' emotional and cognitive biases

This argument leads to a logical follow up conclusion: By striking down collaborations with FCs, platforms potentially enable hate narratives. In other words, if Meta expands its decision to crack down on collaborations with FCs, or if other platforms decide to follow suit, they are arguably contributing to enabling hateful messages.

This conclusion merits, however, one important caveat. To restrict hateful messages, platforms should not only maintain or expand their collaborations with FCs, but remove barriers to checking certain content, such as political content. Currently, platforms such as Meta impose restrictions on the verification of political speech. As a fact-checker pointed out 'The bar to be a political figure [under Meta's fact-checking policies] is very low – just run for a local county council...and you can lie your heart out!'(Gutierrez et al., 2025). Consequently, keeping these limitations in place, platforms enable the spread of hateful populist messages by political leaders.

2.Broader legal and regulatory consequences of the findings

The legal and regulatory consequences of the findings above are twofold. First, from a state perspective, hate speech laws should account for FCs' role in promoting a heterogeneous information environment. As such, States normative frameworks and policies should incentivise FC-platforms collaborations. Second, collaborations between FCs and platforms ensure an approach which is in line with the ECHR. I will assess each issue in turn.

First, as argued, a significant number of papers found links between misinformation and hate speech. From an empirical perspective, Liu et al. (2024) demonstrated that, while there are links between legal regulation and reduced hate speech, there is a non-significant conditional effect of online legal regulation. In other words, the authors argued that legal regulation can constrain hate speech, while also limiting other kinds of lawful speech. Put differently, they claimed that hate speech laws might build an invisible information barrier that operates against the regulation's goals. They further expand on the argument that, 'to eliminate hate speech, people need to be exposed to adequate and heterogeneous information to nurture a tolerant attitude and an open mind' (Liu et al., 2024, p. 542).

Expanding from their argument, hate speech laws must account for FCs' role in ensuring that people are exposed to heterogeneous and reliable information, fostering a tolerant attitude that can, indirectly, help curb hate speech. Instead of enacting hate speech laws that seek to restrict or eliminate certain forms of speech, States can benefit from enacting laws and policies that promote collaborations between FCs and platforms and promoting the role of FCs more broadly speaking.

Second, by collaborating with FCs, platforms ensure an approach that not only observes the DSA and their obligations under Union law, but also the ECHR. Notably, the Article 10 ECHR protects speech that offends, shocks or disturbs, and any measures restricting these kinds of speech must be prescribed by law, in pursuance of a legitimate aim, and be necessary and proportionate. A legal approach which promotes counterspeech instead of restricting speech is preferable as it would not trigger an interference with Article 10 Rights. In other words, by collaborating with FCs, platforms do not interfere or restrict freedom of expression: FCs promote counterspeech instead of censoring users.

In short, overall, the literature supports a normative case for sustaining FC–platform collaborations. While empirical proof of their efficacy remains limited, their alignment with freedom of expression principles and their potential to curb misinformation-induced hate speech justify their continued promotion in human rights frameworks.

Conclusion

In conclusion, this paper considered the implications of collaborations between platforms and FCs to curb online hate and populist narratives. Through a systematic literature review, this study contributed to three scholarly debates. First, I considered how misinformation and hate speech intersect online. Second, I assessed how fact-checking practices may indirectly address hate speech. Finally, I consider the existing evidence on how collaborations between platforms and FCs can reshape responsibility for harmful content under European human rights law.

The paper revealed an important gap in the literature: I found no theoretical or empirical works assessing the impact of collaborations between FCs and platforms in tackling hate speech and populist narratives. However, I found important evidence on broader issues. For instance, there is extensive literature on how misinformation and hate speech intersect online, as well as on the importance of fact-checking initiatives to tackle misinformation and platforms' terms of conduct to restrain hate speech.

Admittedly, some authors emphasise the limitations inherently associated with fact-checking initiatives (such as primarily reaching a different target audience, not necessarily the hate speech spreaders, as argued by Roozenbeek et al. (2023)). Yet, by extrapolating the existing evidence, I argue that FCs can still play a crucial role as a tool to tackle populist hate narratives. This is because, if there is as solid relationship between misinformation and hate speech as portrayed by the literature, these initiatives can still be valuable to constrain hate speech without censorship. Consequently, I argue that FC-platform collaborations advance the fight against hate speech in line with Article 10 ECHR. Future work should, however, consider in greater depth empirical evidence of the efficacy of these collaborations, as well as the consequences of scaling down or withdrawing from such collaborations.

List of papers considered

1 Abdul Reda, A., & Alkhonin, A. (2025). More pressing matters: Can priority reorientation beat online misinformation? *Journal of Computational Social Science*, 8(2).

2 Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2024). Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. *ACM Computing Surveys*, 56(3), 1–17.

3 Asardag, D. (2025). Feminist exploratory interpretive study of the content policy changes of Meta and the corresponding news coverage. Frontiers in Communication, 10.

4 Baptista, J. P., Gradim, A., & Fonseca, D. (2024). Populist Leaders as Gatekeepers: André Ventura Uses News to Legitimize the Discourse. *Journalism and Media*, 5(3), 1329–1347.

5 Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature* (London), 630(8015), 45–53.

6 Caldevilla-Domínguez, D., Barrientos-Báez, A., & Padilla-Castillo, G. (2023). Dilemmas Between Freedom of Speech and Hate Speech: Russophobia on Facebook and Instagram in the Spanish Media. *Politics and Governance*, 11(2), 147–159.

7 Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities & Social Sciences Communications,* 9(1), 1–18.

8 Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, 18.

9 Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, 11(1), 22083–12.

10 Doncel-Martín, I., Catalan-Matamoros, D., & Elías, C. (2023). Corporate social responsibility and public diplomacy as formulas to reduce hate speech on

social media in the fake news era. *Corporate Communications,* 28(2), 340–352.

11 Downing, J., & Dron, R. (2020). Tweeting Grenfell: Discourse and networks in critical constructions of British Muslim social boundaries on social media. *New Media & Society*, 22(3), 449–469.

12 Fraser, R. (2023). *How to talk back: Hate speech, misinformation, and the limits of salience*.

13 Garg, V., Xu, G., & Singh, M. P. (2025). Understanding Inciting Speech as New Malice. *IEEE Transactions on Computational Social System*s, 12(3), 947–956.

14 González-Aguilar, J. M., Segado-Boj, F., & Makhortykh, M. (2023). Populist Right Parties on TikTok: Spectacularization, Personalization, and Hate Speech. *Media and Communication* (Lisboa), 11(2), 232–240.

15 Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy* (University Park, Pa.), 9, 336–369.

16 Komendantova, N., Erokhin, D., & Albano, T. (2023). Misinformation and Its Impact on Contested Policy Issues: The Example of Migration Discourses. *Societies* (Basel, Switzerland), 13(7), 1–16.

17 Kumar, A., & Maurya, M. K. (2024). Online Public Sphere and Threats of Disinformation, Extremism and Hate Speech: Reflections on Threat-Mitigation. *The Journal of Communication Inquiry*.

18 Liu, Z., Luo, C., & Lu, J. (2024). Hate speech in the Internet context: Unpacking the roles of Internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. *Information Developmen*t, 40(4), 533–549.

19 Mayagoitia-Soria, A., González-Aguilar, J. M., Gómez-García, S., & Paz-Rebollo, M. A. (2024). "Drop a Bomb on Them… and Problem Solved!" An Analysis of Poverty Discourse on TikTok. *International Journal of Communication*, 18, 1135–1156.

20 Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, 38(2), 128–146.

21 Mosleh, M., Cole, R., & Rand, D. G. (2024). Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS Nexus*, 3(3), 1–4.

22 Munn, L. (2024). Misinformation's missing human. *Media, Culture & Society,* 46(6), 1287–1298.

23 Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*, 23(6), 1415–

1442.

24 Roberts-Ingleson, E. M., & McCann, W. S. (2023). The Link between Misinformation and Radicalisation: Current Knowledge and Areas for Future Inquiry. *Perspectives on Terrorism* (Lowell), 17(1), 36–49.

25 Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, 28(3), 189–205.

26 Sehat, C. M., Li, R., Nie, P., Prabhakar, T., & Zhang, A. X. (2024). Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1–36.

27 Silva, A. de, & Parker, C. (2025). Platformed hate speech against women: Beyond self-regulation. *UNSW Law Journal*, 48(2), 637–678.

28 Šori, I., & Vehovar, V. (2022). Reported User-Generated Online Hate Speech: The 'Ecosystem', Frames, and Ideologies. *Social Sciences* (Basel), 11(8), 375-.

29 Vasist, P. N., Chatterjee, D., & Krishnan, S. (2024). The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers*, 26(2), 663–688.

30 Vicari, R., Elroy, O., Komendantova, N., & Yosipof, A. (2024). Persistence of misinformation and hate speech over the years: The Manchester Arena bombing. *International Journal of Disaster Risk Reductio*n, 110, 1–15.

31 Vitullo, A. (2021). The Online Intersection among Islamophobia, Populism, and Hate Speech: An Italian Perspective. *Journal of Religion, Media and Digital Culture,* 10(1), 95–114.

32 Zhou, J., Verma, G., Zhang, L., Chang, N., & De Choudhury, M. (2025). Harm in Layers: Compositions of Misinformative Hate in Anti-Asian Speech and Their Impacts on Perceived Harmfulness. *Proceedings of the ACM on Human-Computer Interaction*, 9(2), 1–22.

References

Abdul Reda, A., & Alkhonin, A. (2025). More pressing matters: Can priority reorientation beat online misinformation? Journal of Computational Social Science, 8(2).

Adam, D. (2025). Does fact-checking work? What the science says. Nature (London).

Alkiviadou, N. (2025, March 7). Hate Speech, Positive Obligations And Free Speech: The ECtHR's Expanding Framework In Minasyan And Others V. Armenia (2025). Strasbourg Observers. https://strasbourgobservers.com/2025/03/07/hate-speech-positive-obligations-and-free-speech-the-ecthrs-expanding-framework-in-minasyan-and-others-v-armenia-2025/

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2024). Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. ACM Computing Surveys, 56(3), 1–17.

Asardag, D. (2025). Feminist exploratory interpretive study of the content policy changes of Meta and the corresponding news coverage. Frontiers in Communication, 10.

Bélair-Gagnon, V., Larsen, R., Graves, L., & Westlund, O. (2023). Knowledge Work in Platform Fact-Checking Partnerships. International Journal of Communication, 17, 1169–1189. https://doi.org/1932–8036/20230005

Bengtsson, M., Schousboe, S., Farkas, J., Schjøtt, A., Kjeldsen, J. E., & Hess, A. (2025). Fact-Checkers, Tech-Giants, and Algorithmic Systems: Between Autonomy and Automation in the Relational and Dispersed Construction of Ethos. In Ethos, Technology, and AI in Contemporary Society (1st ed., pp. 249–274). Routledge.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. Nature (London), 630(8015), 45–53.

Cavaliere, P. (2020). From journalistic ethics to fact-checking practices: Defining the standards of content governance in the fight against disinformation. Routledge, Taylor & Francis Group.

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. Philosophy Compass, 18.

CM/Rec(2022)16 - Recommendation of the Committee of Ministers to Member States on Combating Hate Speech (2022).

Erjavec, K., & Kovačič, M. P. (2012). "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. Mass Communication and Society, 15(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Gibson, R. C., Meiklem, R., Moncur, W., & Ruthven, I. (2025). Online Information Disclosure and Information Privacy Practices During Significant Life Transitions: A Scoping Review. CHIIR '25: Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval, 42–56.

González-Aguilar, J. M., Segado-Boj, F., & Makhortykh, M. (2023). Populist Right Parties on TikTok: Spectacularization, Personalization, and Hate Speech. Media and Communication (Lisboa), 11(2), 232–240.

Graves, L. (2017). Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking. Communication, Culture and Critique, 10(3), 518–537. https://doi.org/10.1111/cccr.12163

Gutierrez, N. S. G., Mota, J. da C., & Stremlau, N. (2025, November). ReMeD Policy Brief n1: Building a Stronger Information Ecosystem through Content Moderation: Perspectives from European fact-checkers. ReMeD - Resilient Media for Democracy in the Digital Age. https://resilientmedia.eu/?p=1899

Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. Journal of Information Policy (University Park, Pa.), 9, 336–369.

Hietanen, M., & Eddebo, J. (2023). Towards a Definition of Hate Speech—With a Focus on Online Contexts. The Journal of Communication Inquiry, 47(4), 440–458.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Kahn, G. (2025, January 17). Amid war, vicious attacks and political turmoil, global fact-checkers fear the impact of the end of Meta's programme. Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/news/amid-war-vicious-attacks-and-political-turmoil-global-fact-checkers-fear-impact-end-metas

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

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. BMJ (Clinical Research Ed.), 339, b2700. https://doi.org/10.1136/bmj.b2700

Liu, Z., Luo, C., & Lu, J. (2024). Hate speech in the Internet context: Unpacking the roles of Internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. Information Development, 40(4), 533–549.

López-Borrull, A., & Lopezosa, C. (2025). Mapping the Impact of Generative AI on Disinformation: Insights from a Scoping Review. Publications (Basel), 13(3), 33-.

Mahl, D., Zeng, J., Schäfer, M. S., Egert, F. A., & Oliveira, T. (2024). "We Follow the Disinformation": Conceptualizing and Analyzing Fact-Checking Cultures Across Countries. The International Journal of Press/Politics.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. Television & New Media, 22(2), 205–224.

Mayagoitia-Soria, A., González-Aguilar, J. M., Gómez-García, S., & Paz-Rebollo, M. A. (2024). “Drop a Bomb on Them… and Problem Solved!” An Analysis of Poverty Discourse on TikTok. International Journal of Communication, 18, 1135–1156.

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. Social Science Computer Review, 38(2), 128–146.

Munn, L. (2024). Misinformation’s missing human. Media, Culture & Society, 46(6), 1287–1298.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. Communications of the Association for Information Systems, 37, 43-.

Pentney, K., & Shattock, E. (2025). Disinformation and Democracy on the Docket: Reformulating the Approach to Electoral Disinformation under the ECHR. Oxford Journal of Legal Studies. https://doi.org/10.1093/ojls/gqaf026

Pérez Rastrilla, L., Sapag M., P., & Recio García, A. (2023). Fast Politics: Propaganda in the Age of TikTok. In Fast Politics: Propaganda in the Age of TikTok (1st ed. 2023.). Springer Nature Singapore.

Petticrew, M., & Roberts, H. (2008). Systematic reviews in the social sciences: A practical guide (1st ed.). Wiley.

Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. New Media & Society, 23(6), 1415–1442.

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. European Psychologist, 28(3), 189–205.

Sarah Grevy Gotfredsen. (2023, December 6). Q&A: What happened to academic research on Twitter? Columbia Journalism Review. https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

Schradie, J. (2019). The revolution that wasn’t: How digital activism favors conservatives. In The revolution that wasn’t: How digital activism favors conservatives. Harvard University Press.

Sehat, C. M., Li, R., Nie, P., Prabhakar, T., & Zhang, A. X. (2024). Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), 1–36.

Silva, A. de, & Parker, C. (2025). Platformed hate speech against women: Beyond self-regulation. UNSW Law Journal, 48(2), 637–678.

Westlund, O., Belair-Gagnon, V., Graves, L., Larsen, R., & Steensen, S. (2024). What Is the Problem with Misinformation? Fact-checking as a Sociotechnical and Problem-Solving Practice

Journalism in Exile: Resistance and Vulnerability under Hate Speech

Kezban Karagöz, Utrecht University

The press has always been vital to democracies. The media plays a crucial role in informing the public, shaping elections and enabling citizens to compare political parties. In this respect, the media is considered the "fourth power," following law enforcement and the judiciary (Besley & Prat, 2006). This is also the case in Turkey. Over time, newspapers moved beyond being mere news outlets to become determinants and guides of public opinion and tools in party politics (Habermas, 1990). While ideological hegemony is established (Gramsci, 1971), the media is a key instrument in manufacturing consent (Chomsky, 1995). Although the printing press and the first newspapers arrived later than in Europe, Ottoman and early Republican elites quickly realized the power of the press. Press movements that began in the late Ottoman period influenced the founding of the Republic and later became a strength of the regime (Topuz, 2003). Turkey developed a media system where the press was an indispensable partner of political power. In this environment, the AKP government made the media one of its primary targets. After an economic collapse, the party unexpectedly seized power and, following elections won on promises of secular democracy, gradually consolidated control. Taxes and subsidies were used as instruments against media companies. ATV, one of the most powerful mainstream media groups, was purchased by a businessman close to Erdoğan, and others followed. Over time, large parts of the media yielded. Erdoğan also used media power to polarize society, fragmenting it into opposing camps (Khrimian, 2022).

From Gezi to the Coup Attempt: Authoritarian Turn and Media Capture

The most significant rupture in this trajectory occurred with the Gezi Park protests (Ataman & Çoban, 2019). Plans to transform the park into a shopping mall sparked protests by young people. Harsh police interventions escalated tensions, while the government opted for divisive rhetoric instead of dialogue. Erdoğan stigmatized protesters and deepened social polarization. The elections held after the Gezi protests resulted in a resounding victory for Erdoğan, leaving protesting youth as a marginalized minority. A "Gezi diaspora" emerged as many young people left the country (Zambrana, 2016). Gezi also marked a media turning point. Mainstream channels such as CNNTürk avoided reporting the protests while the government's stance was uncertain, creating a media environment aligned with power. Protesters turned to social media, producing their own activist media and using platforms such as Twitter to broaden the protests and reach the international press in English (Karagöz, 2013; Çoban et al., 2018). This loss of control over the narrative became a milestone for Erdoğan's media strategy. Meanwhile, the corruption revelations implicating the AKP and Erdoğan's family (Lowen, 2014) paved the way for a broader crackdown. Opposition media outlets were shut down en masse, particularly those at the epicentre of tensions after the 17–25 December corruption probes. Media outlets were scapegoated and targeted as part of a larger political strategy. The coup attempt of 15 July 2016 became a milestone in the trajectory of press freedom in Turkey. A total of 178 media outlets were closed by decree between 20 July and 31 December 2016; closure orders were lifted for only nine (Salman & Ergün, 2017). According to RSF, by the end of 2017 some 520 journalists faced the threat of imprisonment because of their work (rsf.org, 2018). Successive declarations of a state of emergency (OHAL) and the suspension of ordinary legal procedures undermined judicial security. Institutions withdrew from the sphere of control: the armed forces, once central to Turkey's political climate, were sidelined and power became concentrated in individuals (ARAS, 2017). The post-coup environment, in which the perception of security was rebuilt, can be seen as an example of internal securitization in the sense used by the Copenhagen School (Aydemir & Decker, 2024). Journalists were suddenly turned into criminals, and safe working conditions became impossible. Many fled to Europe, the United States and Canada; not only famous figures such as Can Dündar (2019), but also reporters and editorial staff (rsf.org, 2017; Culebras, 2020). These dynamics—authoritarian media capture, securitisation, hate speech, troll mobilisation and forced migration—set the stage for the next sections of the chapter, which focus specifically on exiled journalists under the shadow of digital hate speech, their experiences of harassment and their evolving strategies of solidarity and resilience.

Hate Speech in the Shadow of Populist Authoritarianism

Freedom of expression is one of the most important concepts in democracies. It is guaranteed by various international instruments and includes the human right to information, which is vital for forming public opinion and making social choices. At the same time, freedom of expression is not absolute; it is limited when it violates personal rights (Zubčević, Bender & Vojdović, 2017). Article 20 of the ICCPR explicitly prohibits propaganda for war and advocacy of national, racial or religious hatred that incites discrimination, hostility or violence (ICCPR, 1992). In Türkiye, where democratic culture remains fragile, discriminatory and prejudiced news language is widespread, as in many parts of the world. The media is particularly discriminatory towards certain groups and minorities (Güvengez, Saç & Sert, 2019). Nation-state-focused rhetoric has long targeted Armenians, Syriacs, Greeks and Jews, who were among the four groups most exposed to hate speech in a 2019 report by the Hrant Dink Foundation (Güvengez, Saç & Sert, 2019). The assassination of journalist Hrant Dink, repeatedly targeted with hate speech based on his Armenian identity, and the state's reluctance to identify the perpetrators, is emblematic of this pattern—and occurred in a relatively "democratic" period (Freely, 2012).Turkey has historically struggled with democratic processes. Elections have been treated as the primary marker of democracy, yet marked authoritarianism has been visible, particularly after the July 15 coup (Gürkan, 2024). The Gezi protests became a crucial authoritarian experiment. Old labels such as "leftist" or "other" proved inadequate to stigmatize protesters who lacked a classical organisation and mobilized around values rather than rigid ideology. Erdoğan coined the term "çapulcu" (looter) to scapegoat them and render their rights-based struggle invisible (Harding, 2013; Uluğ & Acar, 2018). Following Gezi and later the failed coup attempt, repression of protests became permanent policy. BBC reporter Selin Girit, for instance, was attacked as a "traitor" by the then-mayor of Ankara for her tweets (Gülcan, 2013). Journalists broadcasting individually from protest sites and using personal social media to bypass editorial pressure were sanctioned: many who voiced their views online were fired (Freedom House, 2013). The successive states of emergency after July 2016 and the suspension of legal safeguards sparked debate about a shift from oppressive democracy to single-party dictatorship (İlkiz, 2016). In this climate, hate speech became a central tool of populist authoritarianism, targeting journalists and dissidents as internal enemies. During and after July 2016, institutions assuring the functioning of the state—above all, the legal system—were effectively suspended. The absence or capture of the state facilitated violence against the press, weakened the economic foundations of news organisations and eroded the rule of law. In this context, only media aligned with the regime's rhetoric could survive. The press became not a democratic tool but an instrument for constructing the regime's legitimacy. Many journalists faced a form of "statelessness" in this precarious environment (Waisbord, 2007). For those working in KHK-closed outlets across the political spectrum, forced migration often appeared to be the only solution (Ünsa & Erzurumluoğlu, 2024). Journalists with strong personal brands and large social media followings became especially visible targets (Taşdelen, Ayaz & Coşkun, 2019). Erdoğan uses all the state's authority to ensure the harshest punishments for the journalists he targets. Accusing journalists of being "agents" and "terrorists disguised as journalists," Erdoğan has also pursued journalists who have served time in prison, whose assets have been forcibly seized, and who have been forced to continue their careers abroad due to arrest warrants issued against them. (Ceyhan, 2023) The Istanbul 14th High Criminal Court requested that a red notice be issued for journalist Can Dündar in another case separate from the case file regarding the stopping of MİT trucks. (euronews.com, 2021) Even though Dündar is abroad, he has always been a target of Erdoğan's troll team. (Dündar, 2019) Some were first dismissed from their institutions and then, via social media, turned into entrepreneurial or independent journalists. Others were pushed into exile while continuing to report. BBC (2017) While exploiting the Turkish diaspora abroad for domestic political communication, it also aims to sabotage the security of journalists abroad and ultimately silence them by labeling them agents, traitors, and terrorists. (DW, 2024) The Turkish government has failed to secure the acquittal of various journalists, particularly those in Europe. However, it has placed those who continue to practice journalism on asset freeze lists, labeling them "terrorists" and "terrorist financiers." It has also engaged in a form of data manipulation, to some extent distorting data collected by financial intelligence agencies that provide information to banks.

"This puts targeted individuals in double or even triple jeopardy after they fled the persecution of their own government and were forced to survive and build a new life in a foreign country. Now, they are being targeted once again by private companies for this defamatory information, which violates data protection laws," the SCF warned. (Bozkurt, 2024) This decision, published in the Official Gazette of the Republic of Turkey, labels journalists as "fugitive terrorists." The second step is to exploit the Interpol system to restrict the financial movements of exiled journalists in various countries, effectively rendering them unable to practice their professions. Similarly, the payment flows of journalists broadcasting on YouTube and receiving Patreon support have been sabotaged. (internationaljournalists.org, 2022)

During this period, journalists' addresses abroad were publicized on social media. The journalists received numerous threatening messages during this period, escalating tensions. (internationaljournalists.org, 2021)Furthermore, the regime's FETÖ hate speech, primarily directed at groups aligned with the Gülen movement, has also been used to marginalize journalists from different factions. Exiled journalists from different factions, such as Can Dündar, Cevheri Güven, and Abdullah Bozkurt, are all being labeled as FETÖ fugitives by troll accounts.

Trolls, Digital Authoritarianism and the Stigmatization of Journalists

Many journalists who published stories of public interest based on leaked official reports were prosecuted on charges of "propaganda for a terrorist organisation." Prominent professionals who left mainstream newsrooms became increasingly reliant on alternative and hybrid media platforms that mixed professional and citizen journalism; a smaller group tried to continue individual reporting via personal social media accounts (Ataman, 2018). As repression and arrests intensified, exile journalism became a necessity rather than a choice. Through a phased and strategically coordinated process, Erdoğan's AKP consolidated control over much of the mainstream media. Ownership was reorganised through tenders, tax penalties and political pressure to serve party interests. The Gezi protests, however, demonstrated that social media could enable organically organised publics to bypass mainstream outlets and create powerful counter-discourses. Realising the threat, Erdoğan and his advisors sought to extend control into digital spaces as well, organising them in line with their political discourse (Akiş, 2022). In response to Gezi, the AKP established its first organised “troll army” to undermine criticism and weaken the opposition. Troll teams and “AKbots” were formed to spread propaganda and attack opposition accounts (Yilmaz & Kenes, 2023; Akiş, 2022). A 2023 ACM Web Conference study identified Turkey as one of the most active countries for bot networks on Twitter. These trolls and bot accounts aim to curb emerging public debates by harassing activists, journalists, MPs and opposition politicians. Almost every post of certain “usual suspects” is targeted with insults, accusations of terrorism or treason, and threats of arrest—intended to intimidate and silence dissent (Saka, 2021; Saka & Karataş, 2017). Online harassment is devastating because trolls who write hateful messages have a significant advantage over the journalists they target: virality. This means hateful content spreads very quickly. This creates pressure. Reducing this contagion through methods like media literacy is one suggestion. Urging politicians to take action is another option. (Maarouf, Pröllochs, & Feuerriegel, 2024) However, in countries like Turkey, political figures themselves are spreading this hateful rhetoric.

One common tactic of trolls is to seize control of the social media accounts of journalists or activists and share government propaganda messages through them. The Twitter account of Der Spiegel magazine’s editor-in-chief, Klaus Brinkbäumer, was hacked using a “double switch” tactic on January 14, 2018. The account then shared a photo of Erdoğan and the Turkish flag, along with a message in Turkish. (RSF, 2018.) There are also studies that frame hate speech as extremist discourse. For example, new media criticism and skepticism are common stereotypes among online right-wing supporters in Germany, the US, India, Denmark, Turkey, Hungary, and other countries. Looking at these contexts, we can see that people engaging with digital culture are extremist forms of speech that allow them to say things they wouldn’t say in “real life” interactive situations. Internet memes and trolling are examples of how different manifestations of hate speech can be amplified. (Udupa, 2021) (Pohjonen & Udupa, 2017) Thanks to their structural design and algorithmic prioritization of interaction, social media platforms are seen as effective in the arithmetic escalation of digital hostilities. These platforms mobilize the public, legitimize hostility, and further add digital violence to traditional battlefields. (Orgeret, Mutsvairo, Mirjam de Bruijn, & Moges, 2025) One way hate speech uses social media algorithms to more effectively utilize them is through the use of hashtags. These tags enable hate speech to spread rapidly. In other words, they are now hatetags. (Orgeret, Mutsvairo, Mirjam de Bruijn, & Moges, 2025)

Journalists have been subject to numerous threats in the past. Online harassment can be a precursor to even more brutal and devastating abuse offline. This can be seen in acts of violence against women. The nature of the online harassment journalists face and the types of journalists most likely to be harassed online are important. Today, there are examples of journalists working specifically on data and refuting the regime’s rhetoric being targeted by both digital and physical violence. (Lewis, Zamith, & Coddington, 2020) (unwomen.org, 2024) (rsf.org, 2021)

Digital media has expanded the scope of communication in many ways. One of these is rights advocacy and activism. Hastags, which provide a specific technological cluster, have also become a tool used by activists. A prominent example is the Meetoo movement, (Xiong, Cho, & Boatwright, 2019) along with examples like Blacklivesmatter and Occupaywallstreet. Researchers have characterized these hashtag studies as hashtag activism.

On Twitter, the function of retweets further contributes to hypertextuality. Sharing original posts helps further disseminate messages and their authors. In particular, citizens have been shown to spread digital movements this way (Wonneberger, Hellsten, & Jacobs, 2020) Metin Cihan shared information about trade with Israel, and this news received thousands of retweets on social media.

While digital platforms turn into conflict areas with tags and the dose of hostility increases, complex realities are simplified and polarization is reinforced. These interactions initially addressed Twitter as a platform for dialogue and symbolic struggle. However, over time, the notion of how these platforms have transformed into a space of symbolic struggle has emerged. Hashtags, for example, function as tools of ideological struggle, amplified by the platform's algorithms that foster diasporic mobilization and emotional content. Hate hashtags, on the other hand, transform identity symbols into tools of hostile collective action and digital violence. In this process, the distinction between "us" and "them" in the digital environment can be reinforced, and polarized users are expected to engage in violent reactions. (Orgeret, Mutsvairo, Mirjam de Bruijn, & Moges, 2025) This is similarly evident in the Tigray conflict in Africa. Hashtags such as "FETÖ," "Virus," "Traitor," and "Çapulcu" can reinforce online hostility and shape the narratives of political power. Tags used to organize activist rights-based campaigns can also be used to spread hate speech. These hate hashtags exploit similar algorithmic mechanisms of new media technology to incite polarization and hatred. How these hashtags can become powerful vehicles of hostility in contentious political environments is a matter of considerable debate. (Orgeret, Mutsvairo, Mirjam de Bruijn, & Moges, 2025)

Research

This chapter focuses specifically on the digital hate speech that exiled journalists face once they have reached relative safety. Exiled journalism is not a new phenomenon, but it has gained renewed momentum amid rising authoritarianism and expanding conflict zones (Tetzlaff & Wimschulte, 2024). Some exiled journalists abandon the profession, some work in other fields while producing occasional news, and others continue full-time journalism in exile. The Committee to Protect Journalists (CPJ) reports that its support to exiled journalists increased by 227% over three years (CPJ 2017), reflecting a global press freedom crisis in which journalists from Turkey form a notable group (Westcott, 2023).Forced to leave their home country in search of protection, they systematically become targets of online harassment as they continue reporting. In some cases, this hostility has crossed the line into physical violence. The chapter examines their vulnerability to digital hate speech, the extent of this harassment, and the solidarity networks and digital resilience strategies they build in response. It draws on in-depth interviews with exiled journalists.

Our research is based on seven semi-structured interviews with diaspora journalists from Sweden, Germany, Denmark, and Canada. Six male and one female journalist were interviewed. Exiled female journalists were contacted but were not interviewed because they chose to remain outside of journalism. Interviews lasted 45 to 70 minutes and were conducted in the journalists' native languages or, due to their locations in different cities and countries, in Turkish via WhatsApp or Signal. All interview notes were later transcribed and translated into English. Participants' in-depth accounts of some incidents during the interviews were not disclosed due to confidentiality. Sampling followed a purposeful snowballing technique. We emailed several diaspora journalists known to have experienced digital violence in exile and then contacted other exile journalists who were connected to them. Participants were asked about (a) their journalism careers and experiences working in exile, (b) their connections with local and transnational journalistic and non-journalistic actors, (c) their close relationships with audiences in the diaspora and their homeland, and (d) their evolving roles as exiled journalists and their impressions of journalism, and (e) their experiences of exposure to digital hate speech and their perspectives on the digital diaspora.

Research Findings

Our findings offer important implications for understanding how journalists adapt in the context of war or political prosecution and how journalistic role perceptions and relationships with fellow reporters and audiences mutate when in exile to keep producing public interest content. Moreover, our results also show how, through diasporic journalistic organisations, communities and journalists partake in cultural preservation, identity reaffirmation, and transnational bridge-building, fostering a sense of kinship that transcends geographical boundaries. Their sense of belonging to their new identities through participation in a digital diaspora that differs from their established communities, their sense of security during this process, and their perspectives on the regime's digital violence are explored.

Exiled Journalism Under the Shadow of Digital Violence and the Long Arm of the Regime

Exile does not bring full protection for journalists fleeing authoritarian repression. Even after crossing borders into countries with strong democratic institutions, digital violence and transnational pressure continue to shape their everyday lives and professional practices. Exiled journalists remain visible—often highly visible—in the digital sphere, which becomes both their main tool of survival and their primary site of vulnerability. The same digital platforms that enable them to continue reporting freely also allow the regime to extend its influence beyond national borders, transforming the online space into a transnational battleground of harassment, intimidation, and psychological warfare. In this environment, exile journalism is not merely journalism practiced abroad; it is a form of precarious resistance carried out under conditions of continuous digital surveillance, stigma, algorithmic suppression, and coordinated online hostility. Although exiled journalists rely on digital tools to access sources, publish stories, fund their work (through platforms such as Patreon), and build communities, these tools are simultaneously exploited by the Turkish government and its supporters in a far more organized and strategic manner. As a result, a digital public sphere in exile has emerged—one in which news censored inside Turkey circulates freely—but this sphere also becomes a target for state-led harassment. The campaign to intimidate and silence journalists has led to incidents that threaten journalists' security in the safe countries they visit. Some of the individuals we interviewed were targeted by the regime's trolls, and this continued in the regime's conventional media.

Digital Harassment as a Transnational Weapon

The Turkish government's campaign to intimidate and silence dissident journalists now extends far beyond Turkey's borders. Troll networks, pro-government media outlets, and covert online operations systematically target exiled reporters, reproducing inside host countries the same narrative of treason and terrorism used domestically. The journalists interviewed for this study recounted numerous incidents in which their safety, privacy, and ability to live a normal social life were compromised as a result of digital harassment.

Psychological Warfare in Exile

A prominent example is Levent Kenez, an exiled journalist living in Sweden. Although he previously worked only as an editor in Turkey, his profile rose significantly after he founded the Stockholm Center for Freedom (SCF), a volunteer initiative reporting on human rights violations in Turkey. SCF quickly attracted the attention of the regime. As Kenez shifted to YouTube broadcasting to comment on Turkish politics, the intimidation campaign escalated. Photographs of his daily life in Sweden were taken and circulated online. His home and office locations were posted publicly, accompanied by accusations of treason. Kenez described this campaign as "psychological warfare," designed not only to intimidate him but also to send a clear message to Turkish citizens and other victims of the regime that dissent—even abroad—would be punished.

Kenez's experience illustrates how digital surveillance, doxing, and targeted smear campaigns merge into a form of transnational repression. Even in Sweden, a country known for strong human rights protections, journalists like Kenez become vulnerable to diaspora-based nationalist networks mobilized by Erdoğan's communication strategy. These networks absorb and reproduce the state's "traitor" discourse, turning exiled journalists into pariahs within parts of the Turkish diaspora. This digital hostility also has geopolitical consequences. During NATO negotiations, Erdoğan publicly named individuals—including Kenez—as "terrorists," demanding their extradition. This placed journalists in a situation of extreme uncertainty, where their safety, their legal status, and their professional identity were suddenly entangled with high-stakes international diplomacy. Despite the tense atmosphere, Kenez continued publishing and strengthened his collaboration with international watchdogs such as Nordic Monitor. The solidarity offered by European media and rights organizations became essential to counter the regime's narrative and provide a protective layer against digital harassment.

Diaspora Hostility and Fear

Another journalist, Hasan Cücük, the Danish representative of the shuttered newspaper Zaman, also faced intense digital harassment. As the child of a working-class immigrant family, he embodies both the vulnerabilities of diaspora identity and the political precarity of exile. Due to the autocratic climate in Turkey, he can no longer travel to his homeland—a form of "diaspora exile" in which migration becomes involuntary and permanent. Online threats have heavily affected his social life. He avoids mosques he once attended due to fears of being recognized or followed. Following a wave of digital threats targeting both him and his family, Danish security services placed him under protection, forcing him to relocate temporarily. Despite this, Cücük continues his reporting, especially on the government's use of the "FETÖ" narrative as a catch-all accusation against critics. He maintains strong digital networks with former colleagues who fled Turkey, creating a dispersed but interconnected community of exiled journalists.

Intersecting Vulnerabilities

Hayko Bağdat, a well-known Armenian journalist and activist, left Turkey after sustained death threats. Now based in Germany and co-founder of Özgürüz.org with Can Dündar, he experiences a unique combination of political, ethnic, and digital vulnerability.Unlike others, the threats directed at Bağdat explicitly reference the 1915 genocide and carry distinct racialized and historical undertones. Turkish state officials have continued to issue threats even after his relocation, demonstrating the regime's willingness to project intimidation across borders. For a period, he required a security guard and attended public demonstrations wearing a bulletproof vest. His activism—to support diverse groups suffering injustice in Turkey—intensified after exile, reflecting how digital violence can paradoxically deepen a journalist's political commitment. Bağdat also noted that blocking his social media visibility in Turkey temporarily reduced the volume of threats, underscoring how platform governance decisions directly affect the safety of exiled journalists.

Gendered Digital Violence

Sevinç Özarslan, who began reporting on arbitrary detentions of women, children, and the elderly under the FETÖ label, faces severe gendered harassment. Her stories on strip searches, imprisoned pregnant women, and babies detained with their mothers quickly spread across social media, leading to a surge of hateful and sexually violent messages. These threats are rooted not only in political hostility but also in patriarchal and misogynistic discourse. Even in Germany, where she resides, she continues to receive messages referencing sexual violence. Özarslan maintains strong professional ties with groups in Turkey and experiences a form of "digital diaspora", where the emotional proximity to her homeland is maintained through constant digital exchange—even as it exposes her to ongoing harassment. Research shows that women journalists are targeted online much more frequently than their male colleagues, and these threats are often of a sexual nature. (Oscepa, 2021). It could be argued that women journalists also face more intersectional bullying in this regard. Specifically, the aim here in Turkey is to silence women journalists, who are the voices of vulnerable groups subjected to systematic violence by the government. The exiled woman journalist we interviewed is a key figure, particularly reporting on the injustices faced by families subjected to statutory decrees.

Weaponizing Extradition and Digital Harassment

Another high-profile case is Bülent Keneş, former editor-in-chief of Today's Zaman. Now a scholar and advocate focused on populism, he became a central figure in Turkey's diplomatic narratives when Erdoğan demanded his extradition during Sweden's NATO negotiations. Photographs of Keneş were circulated online with labels such as "terrorist" and "traitor," echoing the regime's official rhetoric.Although Swedish authorities ultimately protected his rights, the ambiguity created during the negotiations illustrates how digital harassment, political labeling, and international diplomacy can reinforce each other. Keneş notes that his advocacy—rather than his journalism per se—continues to make him a target of the regime's long-arm repression.

Discussion and Conclusion

During the autocratic process in Turkey, journalists at risk, who no longer had legal protection, went abroad to relatively safe countries. This forced migration also forced journalists to be involved in a vulnerable migration process. The coup attempt and the seizure of many media institutions by statutory decrees also led to the disappearance of institutional support. Some journalists have abandoned journalism, either out of necessity or perhaps voluntarily. Those we interviewed who continued their work in exile have become subjects of digital hate speech in this process. The regime's complete dismantling of institutions has paved the way for digital violence to be directed at individuals. This systematic hate speech has also narrowed the security of journalists within the Turkish diaspora living in Europe. A female journalist is being intimidated by content of sexual violence through digital threats.

Exiled journalists increasingly face multifaceted threats. The Erdoğan government, in particular, has adopted stringent measures aimed at curbing the active journalistic work conducted from abroad. Social media—an essential space that provides exiled journalists with visibility, opportunities to communicate with their audiences, and, in many cases, economic support—has become a central target of state intervention. The government has entered into various forms of cooperation with global digital platforms such as Google, YouTube, and Twitter/X, now commonly described as new global gatekeepers. As a result, opposition accounts are being suspended without any judicial procedure. Previously, the accounts of exiled journalists had been rendered inaccessible to the Turkish public except through VPN usage, a practice amounting to formalized state censorship. More recently, the Erdoğan administration has taken steps to pressure these same platforms to shut down exile accounts at the international level as well.

Because exiled journalists often remain intensely focused on developments in Turkey while working with extremely small teams, their opportunities to build or sustain local professional networks in host countries are markedly limited. Many choose not to report digital harassment or threats, either due to habituation, a perceived lack of institutional support, or strategic prioritization of their work. Interview participants nonetheless emphasized the unique privilege of being able to speak out against injustices occurring in their country of origin, despite the severe constraints of exile. They articulated a pronounced sense of responsibility toward their colleagues who remain imprisoned in Turkey, a commitment that continues to serve as a primary source of motivation in the face of ongoing digital and physical risks.

Based on the findings of this article, future research should focus more on exiled journalists' solidarity networks and strategies against digital violence. What should journalists' approaches be to effectively advocate for and combat the digital surveillance and censorship of oppressive regimes? What counter-strategies could be used to counter the silencing, closing, and suspension of Patreon accounts of exiled journalists by global digital companies like Google, YouTube, Meta, and Twitter/X? How can global solidarity networks be more effective in legitimizing and empowering journalists' work? It appears that injustices against exiled journalists working in institutions closed by statutory

References

Akiş, F. A. (2022, March 21). *Heinrich Böll Stiftung*. eu.boell.org: https://eu.boell.org/en/2022/03/21/turkeys-troll-networks

Aras, B. (2017). *Turkish Foreign Policy After 15 July.* https://ipc.sabanciuniv.edu/Content/Images/Document/turkish-foreign-policy-after-july-15- 7fc40f/turkish-foreign-policy-after-july-15-7fc40f.pdf

Ataman, B. Ç. (2018). *Profesyonel Gazetecinin Yurttaş Gazetecilikle İmtihanı.* İstanbul: Kafka.

Ataman, B., & Çoban, B. (2019). Turkey: How to deal with threats to journalism. *Transnational Othering – Global Diversities: Media,* (s. 171-190). Sweden: Nordicom.

Ataman, B., & Çoban, B. (2019). Turkey: How to deal with threats to journalism? E. Eide, K. S. Orgeret, & N. M. (Eds.) *Transnational Othering – Global Diversities: Media, extremism and free expression* (s. 171-190). Göteborg: Nordicom.

Aydemir, S. G., & Decker, P. (2024). Turkish nationalism after the 15 July coup attempt: the making of a new political myth. *Journal of Political Ideologies, 30*(3), pp. 1-18. doi:10.1080/13569317.2024.2356539

BBC (2017). https://www.bbc.com/turkce/haberler-turkiye-41119393. *BBC.*

Besley, T., & Prat, A. (2006,, jun). Handcuffs for the Grabbing Hand? Media Capture and Government Accountability. *The American Economic Review, Vol. 96* (No. 3), pp. 720-736.

Chomsky, N. (1995). *Manufacturing Consent.* Newyork: VINT.

CPJ. (2017). *Media in Exile Think Again* https://cpj.org/2017/07/media-in-exile-think-again/

Culebras, I. M. (2020). *opendemocracy.net*. opendemocracy.net: https://www.opendemocracy.net/en/north-africa-west-asia/matter-life-or-death-syrian journalists-have-nowhere-turn/

Çoban, B., Ataman, B., Erduran, Y., & Aydın, U. U. (2018). *Profesyonel Gazetecinin Yurttaş Gazetecilikle İmtihanı.* İstanbul: Epsilon Yayınevi.

Dündar, C. (2019). *X* https://x.com/candundaradasi/status/1131302049377902593

DW (2020). https://www.dw.com/tr/sar%C4%B1-bas%C4%B1n kartlar%C4%B1-iptal-gazeteciler-isyanda/a-52137083

DW (2024). AKP'nin diaspora politikası. *DW,* https://www.dw.com/tr/akp-nin-diaspora politikas%C4%B1-siyasi-hegemonyas%C4%B1n%C4%B1-peki%C5%9Ftirme aray%C4%B1%C5%9F%C4%B1/a-69172399

Finkel, A. (2017, April 25). *CPJ.com*. CPJ: https://cpj.org/tr/2017/04/gonullu-suc-ortagi/

Freedom House. (2013). *Democracy in Crisis: Corruption, Media,and Power in Turkey.* https://freedomhouse.org/sites/default/files/Turkey%20Report%20-%202-3-14.pdf

Freely, M. (2012). Why they killed Hrant Dink. *Index on Censorship*, pp. 15-29.

Gramsci, A. (1971). *Selections from the Prison Notebooks.* New York: International Publishers.

Gülcan, E. (2013). Medyanın dört aylık gezi güncesi, *Bianet,* https://bianet.org/haber/medyanin-dort-aylik-gezi-guncesi-150727

Gürkan, S. (2024). *Democracy under pressure worldwide: 'Türkiye is a textbook example of how countries become autocracies'*, Leiden University, https://www.universiteitleiden.nl/en/news/2024/04/democracy-under-pressure-worldwide-turkey-is-textbook-example-of-how-countries-become-autocracies

Güvengez, S., Saç, E., & Sert, G. (2019). *Hate Speech And Discriminatory Discourse In Media.* Hrant Dink Foundation . Istanbul: Hrant Dink Foundation Publications. https://hrantdink.org/attachments/article/2727/Hate-Speech-and-Discriminatory-Discourse in-Media-2019.pdf

Habermas, J. (1990,). *Kamusallığın Yapısal Dönüşümü.* (T. Bora, & M. Sancar, Çev.) Istanbul: İletişim Yayınları.

Harding, L. (2013). Turkish protesters embrace Erdoğan insult and start 'capuling' craze. *The Guardian*. https://www.theguardian.com/world/2013/jun/10/turkish-protesters capuling-erdogan

HRW (2016). *Türkiye Basınını Susturmak. Hükümetin Eleştirel Gazeteciliğe karşı Derinleşen Saldırısı,* https://www.hrw.org/tr/report/2016/12/15/297442

ICCPR. (1992). *ccprcentre*. ccprcentre.org:

https://ccprcentre.org/files/media/ICCPR_easy_to_read_commentary_WEB.pdf

İlkiz, F. (2016). Darbe, demokrasi, OHAL, *Bianet*. https://bianet.org/yazi/darbe demokrasi-ohal-177115

International journalists (2021). *Sürgün Gazetecilerin Mal Varlıklarına El Konulması,* https://internationaljournalists.org/tr/surgun-gazetecilerin-mal-varliklarina-el-konulmasi kabul-edilemez/

internationaljournalists.org. (2022). internationaljournalists.org: https://internationaljournalists.org/pro-erdogan-daily-targets-yet-another-journalist-in exile-reveals-his-home-address/

Karagöz, K. (2013, Temmuz 25). Yeni Medya Çağında Dönüşen Toplumsal Hareketler ve Dijital Aktivizm Hareketleri. *İletişim ve Diplomasi(1), 131-156.*, s. 131-156.

Khrimian, A. (2022, april 30). *Democratic Backsliding in the Middle East's Former Beacon of Democracy: Polarization and Media Control by Erdoğan and the AKP*. Democratic Erosion

Lewis, S. C., Zamith, R., & Coddington, M. (2020, September 11). Online Harassment and Its Implications for the Journalist–Audience Relationship. *8*(8), s. 1047-1067. doi:10.1080/21670811.2020.1811743

Lowen, M. (2014). *https://www.bbc.com/news/world-europe-30492348*. BBC https://www.bbc.com/news/world-europe-30492348

Maarouf, A., Pröllochs, N., & Feuerriegel, S. (2024). The Virality of Hate Speech on Social Media. *Proceedings of the ACM on Human-Computer Interaction, 8* (CSCW1), p. 1-22. doi:DOI:10.1145/3641025

Orgeret, K. S., Mutsvairo, B., Mirjam de Bruijn, D. T., & Moges, M. A. (2025, Jul 18). Hashtags, Hatetags and social media campaigns in Ethiopia's Tigray conflict. *Information, Communication & Society (iCS)* . doi:https://doi.org/10.1080/1369118X.2025.2533314

Oscar Westlund, R. K., & Orgeret, K. S. (2022). Newsafety: Infrastructures, Practices and Consequences. *Journalism Practice* (s. 1811-1828 |). U.K: Routhledge. doi:https://doi.org/10.1080/17512786.2022.2130818

OSCE Pa (2021). *2021 Report by the OSCEPA Special Representative on Gender Issues highlights surge in violence against women journalists and politicians.*

https://www.oscepa.org/en/news-a-media/press-releases/press-2021/2021-report-by-the osce-pa-special-representative-on-gender-issues-highlights-surge-in-violence-against women-journalists-and-politicians

Pohjonen, M., & Udupa, S. (2017). Extreme Speech Online: An Anthropological Critique of Hate Speech Debates. *International Journal of Communication*.

RSF (2018). *Online Harassment of journalist* . Reporters Without Borders. https://rsf.org/sites/default/files/rsf_report_on_online_harassment.pdf

RSF (2017). *Journalism death throes after six months emergency*, https://rsf.org/en/journalism-death-throes-after six-months-emergency

RSF (2018) *Turkish media in full battle order for Erdoğan reelection*. https://rsf.org/en/turkish-media-full-battle-order-erdo%C4%9Fan-reelection

RSF (2021). Exiled Turkish Journalist attacked in Germany, https://rsf.org/en/exiled-turkish-journalist-attacked-germany

Saka, E. (2021). Networks of Political Trolling in Turkey after the Consolidation of Power Under the Presidency. *Digital Hate. The Global Conjuncture of Extreme Speech* , p. 240-255.

Saka, E., & Karatas, D. (2017). Online political trolling in the context of post-Gezi social media in Turkey. *International Journal of Digital Television, 8* (3), pp 383–401. doi:doi: 10.1386/jdtv.8.3.383_1

Salman, F., & Ergün, A. (2017). Kapatılan Basın Yayın Radyo Televizyon ve Haber Ajansları, *Bianet*. https://bianet.org/haber/kapatilan-basin yayin-radyo-televizyon-ve-haber-ajanslari-182458

Taşdelen, B., Ayaz, F., & Coşkun, G. (2019). Social Media and Personal Branding: Journalists' Preferences For Brand-Shaping Practices. In B. Taşdelen (edt), *New Approaches in Media and Communication.* doi:https://doi.org/10.6035/149

Tetzlaff, S., & Wimschulte, S. (2024). *Exiled Journalist.* Hamburg: Körber-Stiftung. https://koerber-stiftung.de/site/assets/files/43098/koerber

stiftung_publication_exiled_journalist_communities_in_germany-2.pdf

Topuz, H. (2003). *2. Mahmut'tan Holdinglere Türk Basın Tarihi* İstanbul: Remzi Kitapevi.

Udupa, S. (2021). Digital Technology And Extreme Speech. https://peacekeeping.un.org/sites/default/files/digital_technology_and_extreme_speech_u dupa_17_sept_2021.pdf

Uluğ, O. M., & Acar, Y. G. (2018, December 13). 'Names will never hurt us': A qualitative exploration of çapulcu identity through the eyes of Gezi Park protesters. *British Journal of Social Psychology*, pp. 714–729. doi:https://doi.org/10.1111/bjso.12305

Unwomen (2024) *Creating safe digital spaces free of trolls doxing and hate speech*. https://www.unwomen.org/en/articles/explainer/creating-safe-digital-spaces-free-of-trolls doxing-and-hate-speech

Ünsa, F. B., & Erzurumluoglu, B. (2024). The Quest for Justice Against Human Rights Violations. doi:10.13140/RG.2.2.16929.72802

Waisbord, S. (2007). Democratic Journalism and "Statelessness". *Political Communication, 24* (2), pp. 115-129.

Westcott, L. (2023). *CPJ's support to exiled journalists jumped 227% in 3 years, reflecting global press freedom crisis. CPJ.*

Wonneberger, A., Hellsten, I. R., & Jacobs, S. H. (2020, Feb 04). Hashtag activism and the configuration of counterpublics: Dutch animal welfare debates on Twitter. *Information, Communication & Society*, pp. 1694-1711. doi:doi.org/10.1080/1369118X.2020.1720770

Xiong, Y., Cho, M., & Boatwright, B. (2019, March 1). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement. *Public Relations Review*, pp. 10-23. doi:doi.org/10.1016/j.pubrev.2018.10.014

Yilmaz, I., & Kenes, B. (2023). Digital Authoritarianism in Turkish Cyberspace a Study of Deception and Disinformation by the AKP regimes Aktrolls and Akbots *ECPS* https://www.populismstudies.org/digital-authoritarianism-in-turkish-cyberspace-a-study-of deception-and-disinformation-by-the-akp-regimes-aktrolls-and-akbots/

Yilmaz, I., & Kenes, B. (2023). *Digital Authoritarianism in Turkish Cyberspace: A Study of Deception and Disinformation by the AKP Regime's AKtrolls and AKbots*. European Center for Populism Studies (ECPS). https://www.populismstudies.org/digital authoritarianism-in-turkish-cyberspace-a-study-of-deception-and-disinformation-by-the akp-regimes-aktrolls-and-akbots/

Zambrana, M. (2016). The Gezi Diaspora *https://www.equaltimes.org/the-gezi-diaspora?lang=en*.

Zubčević, A. R., Bender, S., & Vojdović., a. J. (2017). *Media regulatory authorities and hate speech*. Strasbourg, France. https://edoc.coe.int/en/media/7431-media-regulatory-authorities-and-hate-speech.html

Part V

Emotions, Psychology, and the Human Experience of Hate

From Label to Harm. How hate speech against individuals associated with the Gülen movement turns "presumptions" into the basis for legal-administrative decisions and leads to real-world harms

Aneta Szarfenberg, The Maria Grzegorzewska University

1. Introduction: What We Discuss and Why Now

This chapter analyses the mechanism by which the labeling of individuals associated (or allegedly associated) with the Gülen movement as "FETÖ" initiates a chain of events leading to real-world harm: from hate speech and dehumanisation, through the treatment of assumptions and associations as evidence, to legal and administrative decisions. The starting point is the Grand Chamber judgment of the European Court of Human Rights (ECtHR) in Yüksel Yalçınkaya v. Türkiye of September 26, 2023, in which the Court found violations of the right to a fair trial, the principle of nullum crimen, and freedom of association. The Court emphasized that convictions based decisively on the mere use of the ByLock application constitute a systemic problem requiring general measures from the state. This judgment highlights the central thesis of this article: association is not proof.

In parallel, research on political and media discourse reveals an entrenched vocabulary of dehumanisation in messages about political opponents, including individuals associated with the Gülen movement. Frequent labels such as "terrorist," "FETÖ," and "virus" are analysed quantitatively in a corpus of Recep Tayyip Erdoğan's speeches (2002-2022), revealing their polarizing function. Press monitoring, such as the Hrant Dink Foundation's Media Watch on Hate Speech and the IPS Communication Foundation/bianet report, document the frequency and function of such terms.

Finally, the quality of evidence and language in the public sphere translate into measurable harm. Qualitative studies on individuals affected by KHK decrees describe a mechanism where the stigma of "terrorist/FETÖ" impacts mental health, leading to isolation and suicidal behaviour. The combination of ECtHR standards, linguistic patterns, and documented harms justifies the remainder of the article, which focuses on demonstrating how to break the chain "from words to harm".

2. The Gülen Movement and the "FETÖ" Label: Conceptual Frameworks

This section outlines the theoretical frameworks underpinning the analysis of the "label-to-harm" mechanism. The analysis rests on two key concepts: Allport's pyramid of hate (Allport, 1954) and an analysis of political rhetoric and its influence on the decision-making process, (Demir, 2025).

Firstly, Allport's pyramid of hate illustrates how lower levels of hostility (stereotyping, denigration) normalise higher levels (discrimination, violence, institutional exclusion). In the context of the Gülen movement, the pejorative labeling of "FETÖ" sets the stage for the treatment of associative indicators as evidence of guilt. It should be noted that while not the central focus, definitions of hate speech provide a crucial backdrop for understanding the escalation mechanism. Using Allport's pyramid can be turned into a practical early-warning dashboard. At levels 1–2 ("language" and "prejudiced acts"), track how often dehumanising labels appear in headlines and TV tickers (Hrant Dink Foundation, 2024), any spikes in "existential threat" headlines (e.g., after major incidents such as 20 Oct 2024), and how far such posts spread on social media. Useful responses here include newsroom style guides that ban dehumanising terms and EU Digital Services Act (DSA) tools: down-ranking such content, adding "speed-bumps" before sharing (e.g., prompts), and prioritising human moderation during spikes. At level 3 ("discrimination"), track the share of cases built only on association—for example, use of the ByLock messaging app—without proof of any criminal act, plus the time to reopen cases and how many convictions are overturned. The fix is a clear ban on "association alone" as proof, fast tracks for retrials/overturning, and administrative reviews that start with reinstating people while the review runs (European Court of Human Rights, 2023). Level 4 ("violence") should be monitored via abuse reports and independent prison monitoring; responses include easy complaint channels and free legal and psychological support (Fidan, 2025).

Secondly, an analysis of political rhetoric, particularly within the speeches of Recep Tayyip Erdoğan, reveals the systematic use of populist and polarizing rhetoric. The employment of dehumanising metaphors ("virus," "traitor," "internal enemy") constructs a narrative of existential threat and facilitates the acceptance of exceptional measures against designated "enemies". This discourse creates a "runway" for the evidentiary shortcuts challenged in the Yalçınkaya judgment. In the context of the Gülen movement, "FETÖ/PDY" has become an entrenched label in official communication and legal documents. This label is not a neutral descriptor; rather, it combines a security frame with the denigration of the opponent, influencing evidence assessment and institutional readiness to impose sanctions. The lack of formal membership in the Gülen movement and the reliance on indirect indicators in assigning affiliations further complicate the situation and contribute to unjust accusations.

3. Method: Desk Research

This analysis employs a desk research methodology, relying on the triangulation of data from diverse sources. A mixed-methods approach, integrating both quantitative and qualitative data, has been adopted to ensure a comprehensive understanding of the phenomenon.

Key sources include: (1) ECtHR Jurisprudence and Commentary: The analysis centers on the Grand Chamber judgment of the European Court of Human Rights (ECtHR) in Yalçınkaya v. Türkiye, along with related commentaries and analyses, providing the legal foundation for the analysis. (2) Hate Speech Monitoring Materials: Data is drawn from reports and analyses by the Hrant Dink Foundation (Media Watch on Hate Speech) and the IPS Communication Foundation/bianet, offering insights into public discourse and its influence on decision-making processes. (3) Quantitative and Qualitative Data: The analysis incorporates quantitative data on the number of dismissals, arrests, and other sanctions, as well as qualitative data from research on the impact of stigmatisation on mental health and family functioning (ECtHR, 2023; Hrant Dink Foundation, 2024a, 2024b; IPS Communication Foundation [bianet], 2024; Fidan, 2025).

Methodological Limitations: The study relies solely on publicly available data, without conducting original fieldwork. Limitations may exist due to incomplete or inconsistent statistical data, and potential biases in qualitative sources. These limitations are mitigated through the triangulation of data from independent sources and the use of cautious language in interpreting the findings.

4. Anchor numbers 2016–2025

Since 2016, the scale of measures targeting people associated with the Gülen movement has been well documented in both governmental and independent sources. According to Minister of Justice Yılmaz Tunç (12 July 2024), in "movement-related" cases more than 700,000 people were subject to proceedings and 13,251 people were in prison—either on remand or convicted (Ministry of Justice of the Republic of Türkiye, 2024a). These measures have persisted despite ECtHR judgments. (Turkish Minute)

Earlier, the U.S. Department of State's 2022 Human Rights Report recorded Turkish authorities' figures of 332,884 detained and 101,000 arrested in the years following 15 July 2016 on terrorism charges linked to the movement. (The same is confirmed in the COI/Home Office review.) (United States Department of State, 2023)

At the same time, the European Court of Human Rights (ECtHR) pointed to the systemic nature of the problem: on the day of the Grand Chamber judgment in Yalçınkaya v. Türkiye (26 September 2023), roughly 8,500 similar applications concerning the right to a fair trial and the nullum crimen principle—where associative evidence such as alleged ByLock use was at issue—were on the Court's docket. The Court classified this as a systemic problem requiring general measures (including a revised approach to evidence allegedly derived from ByLock). (ECtHR, 2023)

Social consequences also include the co-detention of young children with their mothers: ministerial data from October 2024 indicate 706 children living in prisons with their mothers (Ministry of Justice of the Republic of Türkiye, 2024b).

In the academic sector, KHK decrees resulted in 6,081 dismissals of university staff; this figure appears consistently across academic analyses and summary reports (HRFT/Academics for Human Rights, 2018–2020; syntheses 2020–2021). (journals.openedition.org; SAGE Journals)

Internationally, a key marker is the takeover of schools associated with the movement by the Turkish Maarif Foundation: reports from July 2024 (based on TRT Haber) cite 232 schools in 21 countries transferred to Maarif's administration; in previous years the foundation also reported 216 takeovers in 44 countries. (TRT Haber, 2024; Stockholm Center for Freedom, 2024)

These anchor numbers do not replace legal analysis, but they set the scale: the mass character of proceedings, the long tail of cases in Strasbourg, consequences for families and professional groups, and the transfer of education infrastructure abroad.

5. Media and platforms as "accelerators"

Press monitoring. For years, the Hrant Dink Foundation's Media Watch on Hate Speech has systematically tracked media language, maintaining a case archive and periodic reports (e.g., Q1 and Q2 2024). These reviews document how often collective labels ("FETÖ," "traitors," "terrorists") appear and what narrative frames accompany the coverage.

Case studies and numbers. The IPS Communication Foundation / bianet report of 24 July 2024—covering 80 issues of 10 newspapers—found hate speech in 21 of the 66 articles analysed (Hate Speech in Print Media in Turkey). This is a hard indicator of the presence of stigmatizing discourse in the mainstream press. (HDF, 2024a, 2024b); (IPS Communication Foundation [bianet], 2024)

"Spikes" of attention. After the death of Fethullah Gülen (20 Oct 2024), observers noted an intensification of dehumanising content in pro-government media; the Stockholm Center for Freedom's December 2024 report describes this "spike" as a coordinated rhetorical campaign (Stockholm Center for Freedom, 2024). Regardless of how one evaluates that source, it serves as a useful case study showing how high-salience events trigger step-changes in labelling across the media space (Dehumanising a Legacy…, 30 Dec 2024). (stockholmcf.org)

Platforms and regulatory obligations (EU). In the European Union, so-called VLOPs (Very Large Online Platforms) fall under the Digital Services Act: Article 34 requires assessments of systemic risks (including those linked to hate speech and dehumanisation), and Article 35 requires appropriate mitigating measures (e.g., changes to recommendations, "frictions," prioritised human moderation) and periodic reporting. In March 2024, the European Commission issued guidance on recommended measures. These instruments provide operational language for newsrooms and platforms to manage such "spikes." (European Commission; eu-digital-services-act.com)

Evidence of effectiveness. Research on virality and moderation indicates that emotionally charged, antagonistic content has elevated virality rates (analyses on X/Twitter). At the same time, rapid moderation and "frictions" (e.g., confirmation prompts, posting delays) can reduce harm—findings supported by models and literature reviews from 2023–2025. This strengthens the conclusion that anti-dehumanisation standards should cover both editorial language and the parameters of recommendation systems. (dsa-observatory.eu; Electronic Frontier Foundation, European Union, 2022; European Commission, 2024)

Media and platforms are therefore not merely a "mirror" of the conflict—in critical moments they act as amplifiers of the "from person to label" mechanism, which in turn lowers the threshold for presumptions to be accepted as the basis for decisions.

6. Mechanism: from speech to harm

The five-step "from speech to harm" schema is the spine of the argument: it connects the theoretical frame, the empirical findings, and the proposed remedies. It shows how labelling language (the lower tiers of Allport's pyramid) normalises higher levels of hostility (Allport, 1954); within a threat frame, associative presumptions (e.g., use of an app, network ties) begin to substitute for proof, even though the association ≠ evidence standard was confirmed by the Grand Chamber of the European Court of Human Rights in Yalçınkaya v. Türkiye (ECtHR, 2023). Once that shortcut enters practice, it translates into legal-administrative decisions (KHK dismissals, detentions, denial of benefits), producing measurable harms: "civil death," mental-health deterioration, family burdens, and child impacts (Fidan, 2025). Feedback loops then close the circuit: sanctions appear to validate the prior narratives, which in turn lower the threshold for further decisions. At the same time, the schema provides a map of interventions and indicators: newsroom style rules and DSA tools at the language stage; a hard ban on stand-alone association and rapid reviews/retrials at the evidentiary stage and survivor support measures. Read this way, the mechanism below serves not only diagnosis but also early warning and practical correction (Allport, 1954; ECtHR, 2023; Fidan, 2025; HDF 2024a, 2024b; bianet, 2024).

Step 1 — Label and frame. Publicly assigning an "enemy" identity ("FETÖ," "traitors," "virus") installs a dehumanising frame in which "they" are cast as an existential threat. This vocabulary not only polarises but also lowers the institutional threshold for "preventive" action, paving the way for decisions based on associations rather than on case-specific evidence.

Step 2 — Presumptions as a surrogate for proof. Within a threat frame, there is a growing tendency to treat associative indicators (e.g., use of a specific app, network contacts, former workplaces or bank transactions) as proof of guilt. The Yalçınkaya judgment breaks this shortcut: it reiterates that mere "association" (e.g., ByLock) cannot suffice for conviction and cannot replace an assessment of an individual act. This shift matters beyond Turkey—European fair-trial standards and freedom of association assume that correlations and profiles are only operational leads, not evidence of authorship. (ECtHR, 2023)

Step 3 — Translation into legal-administrative decisions. Once presumptions enter practice, a cascade of measures follows: dismissals and professional bans, refusals of passports/benefits, suspension of public services, and—in criminal matters—arrests and convictions partly built on inferences from "associations." This is the point at which language and weak evidence materialise as sanctions in people's lives. (In the media section of the paper, we show how newsroom headlines and TV tickers can legitimise this shortcut.) (Hrant, 2024)

Step 4 — Harms and feedback loops. Qualitative documentation shows that stigma and exclusion after KHK decisions result in isolation, loss of status and income, and—in some cases—depression and increased suicide risk; the stigma spreads to families (Fidan, 2025). These experiences feed further marginalisation ("if they lost their job/have cases, they must be guilty"), closing the loop between language and harm.

Step 5 — Bridge to the European context. In "safe" Western states, an analogous shortcut can be reinforced by algorithmic profiling and network analytics (e.g., in counter-extremism policies or social-control checks). European human-rights bodies warn that such systems must be subject to due-process safeguards (transparency of criteria, contestability, error-rate validation) so that correlations do not morph into pseudo-evidence. This underscores the universality of the speech-to-harm mechanism and the rationale for our recommendations (a ban on stand-alone associative proof, systematic reviews and retrials). (FRA, 2022)

7. Findings

This section presents the main findings of the study, divided into subsections for clarity.

7.1 Language and Normalisation: From Dehumanisation to the Legitimisation of Repression

Analysis of political and media discourse reveals a strong correlation between the use of dehumanising language and the normalisation of repression against alleged members of the Gülen movement. The labeling of individuals associated with the movement as "FETÖ" rarely occurs in isolation from dehumanising metaphors and terms. Frequently encountered terms such as "virus," "traitor," and "internal enemy" are not accidental; they play a crucial role in constructing a narrative of existential threat and legitimizing exceptional measures.

A quantitative analysis of Recep Tayyip Erdoğan's speeches demonstrated the systematic use of polarizing rhetoric that constructs an "internal enemy" and strengthens group cohesion through polarisation. This rhetorical style, consistent with the literature on populism, creates a social climate conducive to simplified assessments and evidentiary shortcuts. Dehumanising language lowers the threshold of tolerance for repressive actions, paving the way for the acceptance of "evidence" based on associations rather than individual actions.

Press monitoring by the Hrant Dink Foundation and bianet confirms the frequent occurrence of dehumanising terms in the media. These terms not only polarise discourse but also contribute to the normalisation of repression. According to Allport's pyramid of hate, lower levels of hate (use of insults, stereotypes) create conditions for the acceptance of higher levels (discrimination, violence). In this context, dehumanising language acts as a catalyst, facilitating the transition from label to sanction. As a result, media and political narratives not only reflect but also reinforce and legitimise repression against the Gülen movement.

7.2 Association as Evidence: The "Evidentiary Shortcut" Mechanism and its Legal Consequences

The findings reveal a systematic use of associative indicators as a basis for assigning guilt in proceedings against alleged members of the Gülen movement. Following the 2016 coup attempt, Turkish authorities widely employed an "evidentiary shortcut," treating connections with the movement as sufficient proof of guilt. These indicators included, among others, alleged use of the ByLock application, network ties, employment history, and small donations. This approach disregarded the principle of nullum crimen sine lege and the right to a fair trial, guaranteed by the European Convention on Human Rights.

The Yalçınkaya v. Türkiye judgment unequivocally condemned this practice, stating that convictions based decisively on associative indicators violate Articles 6, 7, and 11 of the ECtHR. The Court emphasized that correlation cannot substitute for individual proof of guilt, and the use of this "evidentiary shortcut" violates the right to defense and the principle of equality of arms. This judgment has far-reaching practical implications, requiring, among other things, ensuring access to digital data for the accused, replicability of evidence analysis methods, and transparency of error rates. In practice, this means the necessity of applying significantly higher evidentiary standards in cases based on associations with the Gülen movement. Implementation of these standards to date remains contested, which underscores the systemic nature of the problem and the need for wide-ranging reforms in the justice system.

7.3 Gendered harms

While both women and men are targeted, qualitative accounts show gender-specific pathways. Women frequently bear compounded burdens: pregnancy or early motherhood under detention, co-detention with infants, and stigma amplified in community spaces. A stark proxy is the number of children living with their mothers in prison—706 as of late 2024—underscoring how administrative and criminal measures spill over into family life (Fidan, 2025). Men's narratives more often centre on occupational expulsion and criminalisation—dismissals via KHK decrees, prolonged pre-trial detention and the social death that follows a terrorism label—while women additionally describe caretaking strain and community-level shaming. These patterns suggest that gender shapes exposure, setting and secondary effects of the "label →suspicion →sanction" chain, even as core rights violations remain shared (Fidan, 2025).

7.4 Harm to Individuals and Families: Multifaceted Consequences of Repression

Repression against alleged members of the Gülen movement has resulted in a wide range of negative consequences for individuals and their families, extending far beyond the immediate effects of arrests and dismissals.

Dismissals from employment under KHK decrees and labeling as "terrorists" have led to lasting exclusion from the labour market. Job loss meant not only loss of income but also health insurance and other social benefits, as well as difficulties in finding new employment in the private sector due to "blacklisting" and social stigma. This state, referred to as "civil death", is characterized by long-term socio-economic exclusion, deprivation of the ability to function normally, and support a family.

Qualitative research points to serious consequences for mental health. Stigmatisation, social isolation, and job loss lead to depression, anxiety, and, in extreme cases, suicidal thoughts and attempts. Those affected describe feeling stigmatized, avoided, and treated as "monsters", highlighting the profound impact of repression on their self-esteem and identity. Repression is not limited to the accused but also significantly impacts their families. The co-detention of mothers with young children, separation from parents, and stigmatisation at school lead to trauma and mental health disorders in children. These consequences are evident both in statistical data (the number of children in prison with their mothers) and in the accounts of those affected by repression.

Voices from the field (pull-quotes)

"Nobody wants to talk to you; you're like a monster; everyone avoids you." — Participant G-VI.

"People who thought I should be tortured because they called me 'FETÖist' made me live a life of imprisonment in my own home." — Participant G-III (as cited in Fidan, 2025).

"Waking up one morning as a terrorist in a country you've served for years was complete destruction… This was the main factor in my decision to attempt suicide." — Participant G-X.

7.5 Transnational Repression: The Long Reach of Repression Beyond Turkey's Borders

Repression against alleged members of the Gülen movement is not confined to Turkish territory but has a clear transnational dimension. These actions affect the lives of the Turkish diaspora worldwide, and their reach extends beyond Turkey's criminal jurisdiction. The Turkish Maarif Foundation has taken control of numerous schools abroad that were previously associated with the Gülen movement. These takeovers, documented by TRT Haber, exemplify transnational interference in the education system and have far-reaching consequences for Turkish students and teachers abroad. Members of the Turkish diaspora live in constant fear of repression. Many practice self-censorship, avoiding public expression of views that could be interpreted as supportive of the Gülen movement. Others choose to emigrate to countries where they feel safer, representing an additional cost for those affected by repression. There are reports of abductions and forced returns of individuals linked to the Gülen movement from various countries to Turkey. These actions, raising serious concerns under international law, highlight the long reach of Turkish authorities' repression and create additional risks for those living outside Turkey's borders.

8. Conclusion

This article shows that what is often dismissed as "only language" marks the beginning of a predictable process: stigmatising labels create social and institutional tolerance for presumptions, and those presumptions—if left unchecked—become the basis for legal-administrative decisions that cause real harm. This is why it is justified to braid together three strands into a single narrative: the Allport pyramid (lower levels normalise higher ones), the association ≠ evidence standard grounded in the Grand Chamber judgment of the European Court of Human Rights in Yalçınkaya v. Türkiye, and empirical data on health, family and occupational consequences (Allport, 1954; ECtHR, 2023; Fidan, 2025).

The repression of the Gülen movement in Turkey amounts to systemic human-rights concerns, including arbitrary arrests, torture, deprivation of liberty, dismissals from employment, denial of access to healthcare, social stigmatisation, and transnational persecution. The persecution has resulted in severe psychological consequences for victims, including PTSD, depression, suicidal thoughts, the breakdown of family and social ties, and the loss of sense of identity and dignity. These effects are transmitted to subsequent generations. Mass dismissals and stigmatisation have led to mass unemployment, impoverishment, and emigration, causing a "brain drain" and economic damage to Turkey. The dehumanising language used by the Turkish authorities and pro-government media plays a crucial role in legitimizing repression and fostering social acceptance of persecution. This mechanism is well described by Allport's pyramid of hate. Despite numerous reports and international rulings, the international community's response to the repression in Turkey has been limited to date.

A practical conclusion follows from this synthesis: if association is allowed to act as evidence, a feedback loop is created in which anticipatory frames in media and political discourse harden into legal shortcuts, and those shortcuts—once decisions are issued—return to the public sphere as their own legitimation. This is not a mere hypothesis; it is a pattern identified and described both in research on the language of public life (Demir, 2025; Hrant Dink Foundation, 2024) and in case-law and reports on the quality of digital evidence (ECtHR, 2023; Statewatch, 2024).

It is also worth noting that although this case study concerns Turkey and people associated with the Gülen movement, the conclusions are universal. "Safe" Western jurisdictions, too, wrestle with the temptation to use profiling and network analytics in place of traditional evidentiary standards. If human rights are to be more than an abstract declaration, we must keep the evidentiary bar high, distinguish correlation from authorship, and correct language that normalises shortcuts. That is the core of our claim and an invitation to action: implement the Yalçınkaya standard, build pathways to redress, and teach institutions—courts, administrations, newsrooms and platforms—to recognise and stop the 'from words to harm' process before it injures more people.

References

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Bianet / IPS Communication Foundation. (2024, July 24). *Hate Speech in Print Media in Turkey*. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://static.bianet.org/2024/07/bianet-hate-speech-in-print-media-in-turkey.pdf

Demir, V. (2025). Erdoğan's populist rhetoric and hate speech: Anti-opposition discourse and the polarization of Turkish politics. *American Journal of Qualitative Research,* 9(2), 62–78. https://doi.org/10.29333/ajqr/16245

European Commission. (2024, March 25). *Guidelines on recommended measures to Very Large Online Platforms and Search Engines to mitigate systemic risks online* (DSA, Article 35). https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

European Commission. (2024). *Guidelines for providers of VLOPs and VLOSEs on the mitigation of systemic risks for electoral processes.* https://digital-strategy.ec.europa.eu/en/library/guidelines-providers-vlops-and-vloses-mitigation-systemic-risks-electoral-processes

European Court of Human Rights [ECtHR]. (2023). *Yüksel Yalçınkaya v. Türkiye* (No. 15669/20), Grand Chamber judgment of 26 September 2023. https://hudoc.echr.coe.int/eng?i=001-227636

European Court of Human Rights [ECtHR]. (2023, September 26). *Judgment concerning Türkiye* (press release). https://www.echr.coe.int/w/grand-chamber-judgment-concerning-turkiye

European Union. (2022). *Regulation (EU) 2022/2065 on a Single Market for Digital Services* (Digital Services Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

Turkut E. (2024), *ByLock Prosecutions and the Right to Fair Trial in Turkey: The ECtHR Grand Chamber's Ruling in Yüksel Yalçınkaya v. Türkiye, Statewatch*, SSRN

Fidan, F. Z. (2025). Victims of decree laws in Turkey: The relationship between stigma and suicide. *Addiction Research*, 9(1), 1–7. https://doi.org/10.33425/2639-8451.1049

FRA, *Bias in Algorithms – Artificial Intelligence and Discrimination*, Vienna 2022. fra.europa.euRM.coe.int

Hrant Dink Foundation. (2024). *Media Watch on Hate Speech – 2024 selections (Q1 & Q3)*. https://hrantdink.org/en/asulis/activities/projects/media-watch-on-hate-speech

Home Office (UK). (2025, August). *Country Policy and Information Note: Gülenist movement, Turkey*. https://www.gov.uk/government/publications/turkey-country-policy-and-information-notes/country-policy-and-information-note-gulenist-movement-turkey-february-2022-accessible-version

Ministry of Justice of Türkiye. (2024, October). *706 children living in prisons with their mothers, minister says (reported*). https://www.turkishminute.com/2024/10/11/706-children-accompany-mother-in-turkish-prisons-justice-minister/

Statewatch. (2024, March 13). *ByLock prosecutions and the right to fair trial in Turkey: The ECtHR Grand Chamber's ruling in Yüksel Yalçınkaya v. Türkiye*. https://www.statewatch.org/publications/reports-and-books/bylock-prosecutions-and-the-right-to-fair-trial-in-turkey-the-ecthr-grand-chamber-s-ruling-in-yuksel-yalcinkaya-v-turkiye/

Stockholm Center for Freedom. (2024, December 30). *Dehumanizing a Legacy: How Fethullah Gülen's Death Triggered a Campaign of Hate Speech in Erdoğan's Turkey.* https://stockholmcf.org/new-report-exposes-systematic-use-of-hate-speech-following-fethullah-gulens-death-to-dehumanize-his-movement/

TRT Haber. (2024, July). *Turkish Maarif Foundation has taken over 232 schools in 21 countries.* https://www.trthaber.com/ https://stockholmcf.org/turkish-govt-seized-232-gulen-linked-schools-in-21-countries-since-coup-attempt/

Turkish Minute. (2024, July 16). *Turkish gov't seized 232 Gülen-linked schools in 21 countries since coup attempt*. https://www.turkishminute.com/2024/07/16/turkish-govt-seized-232-gulen-linked-school-21-countries-since-coup-attempt/

U.S. Department of State. (2023). *2023 Country Reports on Human Rights Practices: Turkey*. https://www.state.gov/reports/2023-country-reports-on-human-

Digital platforms as Tools for Influence and Polarised Politics

Soraya Afzali, Trinity College Dublin

Introduction

Populism, nationalism, and neo-liberalism are part of contemporary politics in Ireland. Nationalism while ingrained in the fabric of political history of Ireland is not entangled with the accelerating neo-liberal order of the country through populist strategies and rhetorics. It is no longer a sustainable argument to present Ireland as an exception to the realities of populist political strategy seen in other parts of Europe. Over the past two decades, the visibility of new migrant communities has become entangled with the tension of populist nationalism, where digital communication both reflects and amplifies radicalised and hateful narratives. The centre-right, while promoting neo-liberal economic structures in the country, has utilized the rhetoric of 'Irish Identity' politics to gain influence among voters, echoing the postcolonial myth of Ireland as a small, homogenised country, erasing its rise towards a multicultural reality. This is framed by the subtext of Irish identity being 'predefined', where immigrants are only conditionally accepted.

This chapter argues that the recent cases of racism and hate incidents towards the Indian population are the result of a populist nationalism built on the economic foundation of neo-liberalism and mobilised through online platforms. The Indian population in Ireland, who are predominantly professionals, have become the current subject of party politics, where constructed anxieties around Irish identity are defended and promoted indirectly by centre-right, right, and far-right political parties, referred to here as far-right in this paper. This adaptation of the term is because of the usage of far-right narratives in soft an extreme forms in political parties associated with right. The values expressed by far-right political parties, and adopted by the right more broadly, are now deeply entwined with hostility towards the non-Irish population in general and, as recent incidents show, have escalated against the Indian population.

This analysis critically engages with the logic of neo-liberalism and its commodification of mobility, where economic migrants, including those from India, are praised for their economic contribution but are denied belonging and recognition and are framed as a threat to national cohesion. These are factors that contribute to the formation of populist nationalism, a rhetoric promoted by right-wing parties, including the centre-right, right, and extreme right. The promotion of anxieties centred around myth-based narratives in the public discourse permeates online platforms, where a narrow understanding of "values" is used to other immigrants. This article contributes to an understanding of neo-liberalism and its role in the propagation of hate, including racism towards the Indian population, among other economic migrants, and positions not only far-right politics as its driver, but firmly roots the phenomenon in the mainstream centre-right.

Populism, Historical Nationalism: the production of exclusion in Ireland

Populism is built on a binary approach to politics (Mudde 2007; Moffitt 2016; Müller 2017). As a strategy, it relies on left-right, pro-against, and is best explained by the us-them dichotomy. Its rationale is to speak for 'the people' against those positioned as elite, corrupt, or external, which is also emblematic of a binary strategy in politics where exclusion is pivotal. The right-left divide can be traced to the French Revolution, when Jacobins and Girondins sat on opposite sides of the president of the parliament (Mudde 2019). Populism thus lends itself to a polarised duality of right and left, and explains its adaptability to contexts based on steering group membership in one direction at the cost of exclusion of the other (Kirby 2024). These exclusions are understood under the pretext of Bobbio's theory of political distinction, where the level of acceptance of inequality (Lindqvist 2020) can be utilised in determining values. Of course, a multi-party system is also a foundational feature of any democracy; however, in the contemporary politics of Europe, and Ireland specifically, the relationship between immigration and othering has evolved into normative instrumentalization, with rising racism as its consequence.

In Ireland, the relatively long history of fascism and violent party opposition as explored by Pádraige Óg Ó Ruairg (2025) can be traced to the antisemitism of the Irish political elite in the 1920s, the IRA's support of Nazi Germany in the 1940s, and the Ultra-Conservative and Catholic movements of the Blueshirts, influenced by Italian Fascism. These accounts bring together the long history of violence and hatred based on a populist, divisive rhetoric, where such values can, in one way, find their connection to the rise of the far-right in the late 2010s and onward, leading to the riots in November 2023. In November of 2023, groups of far-right agitators mobilised through online channels that caused hours-long riots in Dublin city centre. The riot was triggered by a man stabbing three young children and a care assistant. This included arson attacks and destruction of roads, buses, shops, and tram lines with damages reaching €5 million. The riot was later investigated and connected to far-right mobilisation through WhatsApp and Telegram groups. Pádraige Óg Ó's deep dive into historical events and their pivotal consequences is significant for contextualising current far-right ideologies in Ireland that target immigrant communities. Although targeting immigrants by the far-right in Ireland can be seen as a new development, from a policy perspective, it has evolved out of populist politics and the 2004 Irish Citizenship Referendum that inserted the language of differentiation between 'Irish' and 'non-Irish' into public discourse. The development of far-right ideology popularised through anti-immigrant rhetoric in Ireland from 2018 to 2024 is investigated coherently in The far-right Rise, activities, and international links (2025), edited by Yasmin Ahmed, highlighting transnational influences, particularly through online platforms.

Today, political parties reflect Ireland's historical upheavals. The two leading parties on the centre and centre-right, Fianna Fáil (FF) and Fine Gael (FG), emerged out of division over the Anglo-Irish Treaty of 1921. The Anglo-Irish Treaty of 1921 was an agreement signed between the British government and the representatives of the Irish Republic that concluded the Irish War of Independence. The treaty was also the beginning of the party division among the Irish party politics marking the division of the 1920s Sinn Féin into pro-treaty and anti-treaty parties. The nature of this historical and political division places nationalism at the core of Irish politics. It surfaces as a central element existing in all contemporary political parties, including the centre-left Sinn Féin. Immigration may have become a political issue from 1996 onwards, with immigration peaking in 2002, but the far-right ideologies that quietly underpin Irish party politics can be traced to Ireland's historical roots through the "hegemony of populist nationalist mainstream" (Garner 2007). This soft but nonetheless potent populist nationalism functions as the backdrop of Irish politics in general and has now manifested in the real world with the rise of vocal nationalist and far-right activities, including regular arson attacks on International Protection Accommodations, Direct Provision Centres, among other anti-immigrant protests from 2018 onwards. The target of these far-right activities are perceived asylum seekers or those in Direct Provision Centres or International Protection Applicants who make up the smallest number of the non-Irish population compared to economic migrants, referred here to those residing in Ireland based on work-permits. Although this targeting is specific in its violence, such as the arson attack in October 2025 that risked the lives of children and babies, the effects are widespread and pernicious, creating a sense of alienation among the entire non-Irish population, including Indians who have arrived in Ireland on work permits similar to European migrants. Those seeking work permits, despite sharing an economic migration background with these European economic migrants, have a different experience both bureaucratically and in their daily lives. For instance, non-European migrants require a work permit issued by the Ministry of Enterprise, Trade, and Employment based on a contract offered before applying for an Irish visa from home country. Populist nationalism, normalised in Irish politics, has made this group of economic migrants, perceived as non-European, the target of daily racism in Ireland.

Economic development and the Indian population

From the Celtic Tiger boom of the 2000s onwards, with a brief economic slowdown in 2008-2010, Ireland has enjoyed a lengthy period of economic growth, particularly through tech exports and the pharmaceutical industries. Ireland was of the fastest-growing GDPs in Europe in 2021 and keeping an above average GDP (OECD 2025). The decline of unemployment has been consistent since 2011 onwards (CSO 2024). Nevertheless, levels of disparities still exist in the labour market, best understood through a neo-liberal lens. The lack of class recognition in the real world and in Irish politics is a strong indicator of how neoliberal politics has played out which, best understood through a 'putative middle class' or the 'squeezed middle class' (Meade and Kiely 2020). The principles of neo-liberalism, such as free market, privatisation, and deregulation, with a focus on strong private property rights, are the core underpinning of how politics function (Hathaway 2020; Nofal 2023). Ireland is no exception in this. The inflow of economic migrants to Ireland is embedded in these broader neoliberal dynamics.

A total of 149,200 immigrants entered Ireland in the 12 months leading to April 2024, a record high over the last 16 years in Ireland, this number included the returning Irish people. Based on the Central Statistics Office census of 2022, there are 45,449 Indian nationals in Ireland (CSO 2025). In the same year, close to 10,000 Indian nationals came to Ireland, taking up residence, making them the largest non-Irish group in the country (CSO 2025). In 2024, the highest number of work permits were issued to those coming from India. The numbers are expected to be the same or higher in 2025. Because housing is limited in Ireland, particularly in Dublin—a crisis going as far back as 1980s—means the incoming economic migrants have been spread across Ireland. Incidents of hate directed at Indians have thus spread beyond Dublin, a complex phenomenon involving manipulation of anti-immigrant narratives in less concentrated areas in Ireland, discussed later. This spread of economic migrants across Ireland is also an outcome of the neo-liberal systems of privatisation and property management. This interconnection is thoroughly explained by Raquel Rolnik in her book, Urban Warfare: Housing under the Empire of Finance, which was written in her capacity as the United Nations special rapporteur on adequate housing. While she speaks about the UK housing situation and its government hold, she provides a wider global outlook on housing policies worldwide and its failures within a neo-liberal system. This interconnection also in Ireland explains how houses are not built to meet social needs anymore but to offer investment opportunities in central Dublin and across Ireland. This development has led to a lack of affordable housing for those who are not economically squeezed, and they will have to choose to live in the periphery.

Hate crimes and incidents

The arc of Irish history bends toward the populism that politically divide and the economic advance and inflow of immigration into the country complements these old strategies that adopts to the contemporary environments. The divisions are enforced by the rise of hatred against the Indian population in Ireland through compartmentalisation of immigrant groups based on differences that are racially motivated. These compartmentalisations bypass the economic contribution due to the neo-liberal order and politicises belonging in a nationalistic framework that manifest violence in different forms. The number of hate crimes and incidents recorded by Gardaí in 2024 were 676, including 264 incidents 'anti-race crimes/incidents', the highest number since the system was introduced. These numbers do not cover the extent of 'anti-race incidents' in Ireland, as most are not categorised under either anti-race 'incident' or crime, due to Ireland not having had effective hate crime legislation prior 2024. The 1989 Prohibition of Incitement to Hatred Act is the main legislation in Ireland that addresses hate speech and prohibits the use of language and behaviour that are likely to stir up hatred and the dissemination of racist ideas. In 2024, a new Criminal Justice (Hate Offences) Act was introduced to provide a framework for hate crimes, including hate speech. The An Garda Síochána used a working hate crime and hate incident definition introduced in October 2019 that categorises criminal offences and non-criminal offences (INAR 2019). Apart from legislative barriers, most hate incidents, including racism against marginalised communities such as the Indian population, are hidden hate incidents because not all are reported and recorded, as mentioned in some of the instances below.

Despite this lack of reporting, racially motivated hate-incidents against the Indian population in Ireland rose substantially in 2025. News outlets such as the Irish Independent, Irish Examiner, Irish Times, Stand.ie, among others, covered incidents that included severe physical assaults. The throwing of glass at nurse Jibby Palatty working in one of Dublin's hospitals (Pollak and Gallagher 2025), the physical assault of Indian man working in a tech company in Tallaght on June 19th (Chowdhury 2025), the physical assault of an Indian man working in another tech company resulting in a face fracture in Letterkenny (Nath 2025), the harassment of a 6-year old child in front of her house in south-east Ireland (Stand.ie 2025), and the physical assault of an Indian taxi driver in Ballymun (Walsh 2025) are just a few of the recorded incidents in May-August, 2025 targeting the Indian community. The Christian Science Monitor, citing the Ireland India Council, an independent institution, reported 11 racially motivated attacks against the Indian community from July-August alone. Civil society organisations such as Doras, Immigrant Councils of Ireland, and the Irish Refugee Council expressed concerns about this rise of racism against Indians, including a statement by the Indian embassy in Dublin, political parties, and the president.

The involvement of children and adolescents as perpetrators in some of these assaults is particularly alarming, as it underscores how intolerance has found its way into neighbourhood cultures and even internalised in family settings. These events are also connected to the apparent misinformation spread through social media platforms, including X, Facebook, TikTok, and Telegram. The two cases in Tallaght and Letterkenny were connected to rumours being circulated on these platforms, some with thousands of views. In two cases, these rumours targeted innocent victims, accusing them of inappropriate behaviour towards a child, which were debunked by the victim himself and the Gardaí. Throughout 2025, rumours spread about physical assaults, inappropriate behaviour towards children, and rape, echoing past myth-based moral panics around protecting 'Irish women', 'Irish Nationality', 'Irish jobs', and 'Irish culture'. Myth-based narratives at times of crises—real or imagined—are potent to gain momentum, specifically within a populist rhetorics (Mudde 2007, 2016). Myth-based narratives and framings simplify the nuanced issues, namely the root causes of crises in the first place. The online spread of these narratives through social media is one of the major mechanisms of mobilisation for offline hostility. This pattern echoes far-right ideology through a strategy of online mobilisation, which was recognised and harnessed during the COVID-19 pandemic.

Covid and Far-right mobilisation online

The surge in hate crimes can also be connected to conspiracies that began with QAnon, the extreme far-right movement in the U.S., which reached many nebulous Irish far-right individuals and groups during the anti-lockdown protests (Fattibene et al. 2024). These were often tailored to the Irish context, a narrative flexibility that features prominently among populists and far-right ideologues (Pirro 2024), akin to the change of signifiers that strategises polarity. The COVID-19 pandemic left a few lasting impacts on far-right ideology and its followers. The first was the connective thread of anti-vaccine tendencies as a source of identification and political belonging (Paoletti et al. 2024). Being distrustful of vaccines and the instrumentalization of social media for the dissemination of misinformation prevalent in the far right, was characterized as opposition to a 'corrupt elite'. The anti-institutional stance went beyond health concerns explained by group identification in social psychology (Magnus 2022) where in-group favouritism took place. The political stance of the right wing reflected the 'us' and 'them' rhetoric associated with populism, and the weaponization of conspiratorial fears and anxieties to reinforce in-group identity(Magnus 2022), characterized by resistance to public health measures (Backhaus et al. 2023) (Paoletti et al. 2024). This may have been a feature of far-right ideology before the pandemic but the crisis paved the way for a more concerning change to the political atmosphere: the realization of the power of new media tools. The pandemic enhanced social media platforms as spaces where people could mobilise (Daphi et al. 2024) and facilitated transnational ideological cohesion. This mobilisation paved the way for the misuse of information, especially by radical right-wing parties (Törnberg and Chueri 2025). Misinformation is most potent in times of crises, as was the case with the 'refugee crises' the continuous climate change crisis, and in the case of the pandemic, produced an enduring threat to democracy (Daphi et al. 2024:16). One example of this polarising antagonism can be seen in the arguments put forward by the far right in recent years, specifically during the COVID-19 pandemic, which displayed a polarisation among the people, between those who are pro-vaccine and those who are not. This suggested a type of group membership, exhibited, and eventually internalised. It was in the intersection of these processes that political parties also capitalised on these proposed values for followers.

Conclusion

The pattern of instrumentalisation of immigration and immigrant identities is a recognised strategy among far-right parties. In the Irish context, although and when direct political influence remains limited, this tendency becomes more pronounced during the electoral period, when competition for power amplifies exclusionary narratives. This was the case with the deputy prime minister of Ireland (Tánaiste), Simon Harris, who expressed appreciation for the contributions of Indian immigrants when the Indian embassy issued a statement warning its residents in Ireland about safety measures. A couple of days after the presidential election that took place on 24th of October, with his party's candidate losing the election, he stated : "Our migration numbers are too high, and I think that is really an issue that needs to be considered in a very serious way by the Government. One of the reasons I think they are so high is that there are too many people who come to this country and are told they do not have a right to be here, and it is taking too long for them to leave the country." (O'Toole 2025). Harris is known as a centre-right politician representing Fine Gael. This was after Catherine Connoly's win, a left-leaning candidate for the presidency. Ó Ríordáin of the labour party did call out on Harris's spread of misinformation on immigration (Ó Ríordáin 2025). The online platforms and the ease of communication distribution and mobilisation it provides is only a pivote of these political dynamics in the contemporary politics that has roots in the neoliberal systems.

References

Ahmed, Yasmine. 2025. *The Far-Right in Ireland: Rise, Activities and International Links*. S.l.: Bristol University Press.

Backhaus, Insa, Hanno Hoven, and Ichiro Kawachi. 2023. 'Far-Right Political Ideology and COVID-19 Vaccine Hesitancy: Multilevel Analysis of 21 European Countries'. *Social Science & Medicine* 335:116227. doi:10.1016/j.socscimed.2023.116227.

Chowdhury, Arpita. 2025. 'The Racist Attack on an Indian Man in Tallaght Recently Was Not a Once-off. Ireland Must Act'. *Irish Examiner*.

CSO. 2024. *Labour Force Survey Quarter 3 2024*. Ireland. https://www.cso.ie/en/releasesandpublications/ep/p-lfs/labourforcesurveyquarter32024/unemployment/.

CSO. 2025. *Census of Population 2022 Profile 5 - Diversity, Migration, Ethnicity,* *Irish Travellers & Religion*. https://www.cso.ie/en/releasesandpublications/ep/p-cpp5/censusofpopulation2022profile5-diversitymigrationethnicityirishtravellersreligion/citizenship/.

Daphi, Priska, Cristina Flesher Fominaya, and Eduardo Romanos. 2024. ‘Introduction: Mobilizing during COVID-19: Social Movements in Times of Crisis’. *Social Movement Studies* 23(6):667–75. doi:10.1080/14742837.2024.2406697.

Fattibene, Gabriella, James Windle, Orla Lynch, Grant Helm, Joe Purvis, and Liina Seppa. 2024. ‘The Online Exchange of Conspiracy Theories within an Irish Extreme Right Wing Telegram Group during the COVID-19 Pandemic’. *Routledge* (Behavioral Sciences of Terrorism and Political Aggression):1–19. doi:10.1080/19434472.2024.2409185.

Garner, Steve. 2007. ‘Ireland and Immigration: Explaining the Absence of the Far Right’. *Patterns of Prejudice* 41(2):109–30. doi:10.1080/00313220701265486.

Hathaway, Terry. 2020. ‘Neoliberalism as Corporate Power’. *Competition & Change* 24(3–4):315–37. doi:10.1177/1024529420910382.

INAR. 2019. ‘Why Hate Crime Legislation?’ https://inar.ie/hate-crime-legislation/.

Kirby, Peadar. 2024. ‘The Rise of the Far-Right: From Condemnation to Understanding’. *Project MUSE* 113(449):105–14.

Lindqvist, Jesper. 2020. ‘The Inequalities That Divide – A Theory of Left-Right Politics’. UCD School of Politics and International Relations, Dublin.

Magnus, Kathleen D. 2022. ‘Right-Wing Populism, Social Identity Theory, and Resistance to Public Health Measures During the COVID-19 Pandemic’. *International Journal of Public Health* 67:1604812. doi:10.3389/ijph.2022.1604812.

Meade, Rosie R., and Elizabeth Kiely. 2020. ‘(Neo)Liberal Populism and Ireland’s “Squeezed Middle”’. *Race & Class* 61(4):29–49. doi:10.1177/0306396820905729.

Moffitt, Benjamin. 2016. *The Global Rise of Populism: Performance, Political Style, and Representation*. Stanford University Press.

Mudde, Cas. 2007. *Populist Radical Right Parties in Europe*. Cambridge University Press.

Mudde, Cas. 2019. *The Far Right Today*. Polity Press.

Müller, Jan-Werner. 2017. *What Is Populism?* Penguin Random House UK.

Nath, Sanstuti. 2025. ‘Indian-Origin Man’s Face Fractured In “Unprovoked Racist” Attack In Ireland’. https://www.ndtv.com/world-news/indian-origin-man-left-with-face-fracture-after-unprovoked-racist-attack-in-ireland-8990124.

Nofal, Sulafa. 2023. ‘The Historical Roots of Neoliberalism: Origin and Meaning'. *Brazilian Journal of Political Economy* 43(3):576–91. doi:10.1590/0101-31572023-3444.

Ó Ríordáin, Aodhán. 2025. 'Ó Ríordáin Calls out Tánaiste's Reckless Remarks on Immigration'. https://labour.ie/news/2025/10/30/o-riordain-calls-out-tanaistes-reckless-remarks-on-immigration/.

Ó Ruairg, Pádraig Óg. 2025. *Burn Them Out!: A History of Fascism and the Far Right in Ireland*. Head of Zeus.

OECD. 2025. *OECD Economic Outlook, Volume 2025 Issue 1: Preliminary Version*. Vol. 2025. OECD Economic Outlook. OECD Publishing.

O'Toole, Fintan. n.d. 'Simon Harris Is Deliberately Spreading Disinformation on Immigration'. *Irish Times*.

Paoletti, Giordano, Lorenzo Dall'Amico, Kyriaki Kalimeri, Jacopo Lenti, Yelena Mejova, Daniela Paolotti, Michele Starnini, and Michele Tizzani. 2024. 'Political Context of the European Vaccine Debate on Twitter'. *Scientific Reports* 14(1):4397. doi:10.1038/s41598-024-54863-7.

Pirro, Andrea L. P. 2024. 'The Contemporary Far Right from *Contra* to Control'. *Political Communication* 41(6):1017–22. doi:10.1080/10584609.2024.2414256.

Pollak, Sorcha, and Conor Gallagher. 2025. '"He Kept Saying: What Wrong Have I Done? Why Me?" An Indian Man Is Left Stripped and Bloodied on an Irish Street'. *Irish Times*.

Stand.ie. 2025. 'When Hate Targets the Young: Racist Violence Against Indians in Ireland'.

Törnberg, Petter, and Juliana Chueri. 2025. 'When Do Parties Lie? Misinformation and Radical-Right Populism Across 26 Countries'. *The International Journal of Press/Politics* 19401612241311886. doi:10.1177/19401612241311886.

Walsh, Louise. 2025. 'Indian Taxi Driver Faces Loss of Livelihood as He Recovers from Unprovoked Dublin Attack'. *Irish Times*.

9 789076 566009