

Hate Speech, Information Isolation, and the Making of “Invisible Crime”



solidarity with
OTHERS

October 2026

This report has been published by Solidarity With OTHERS. The views expressed herein are those of the author and do not necessarily reflect the views of any donors or affiliated institutions.

Hate Speech, Information Isolation, and the Making of “Invisible Crime”

Author: Hongyu PAN

© 2026 **Solidarity With OTHERS**

Weiveldlaan 41 Unit D, 1930 Zaventem BELGIUM
www.solidaritywithothers.com
info@solidaritywithothers.com

Solidarity With OTHERS, an international human rights organisation registered in accordance with the Belgian Code of Companies and Associations (**OTHERS AISBL**, BCE number: 0715.742.907).

This publication may be used for advocacy, education, and research purposes, provided the source is acknowledged.

Table of Contents

- 1.** Introduction 2
- 2.** Hate Speech within the Framework of International Human Rights Law 5
- 3.** Structural Changes in Hate Speech in the Digital Age..... 8
- 4.** The Mechanisms Behind Hate Speech 11
- 5.** In-Depth Analysis: Structural Mechanisms of Hate Speech, Information Segregation, and “Invisible Crime” 14
 - 5.1. From Single Acts to Cumulative Harm 14
 - 5.2. Digital Amplification and Information Isolation 15
 - 5.3. Structural Violence and Invisible Crime 15
 - 5.4. European Illustrations..... 16
 - 5.5. Multi-Level Effects of Invisible Crime..... 17
 - 5.6. Risks of Regulatory Overreach 18
- 6.** Policy Recommendations 19
- 7.** Conclusion..... 21
- References 22

1. Introduction

In the era of globalization and digital communication, hate speech has evolved from a marginal phenomenon in traditional public spaces into a global human rights issue characterized by transnational dissemination. Digital platforms allow hate speech to spread across borders without geographic limits. Algorithms and online communities amplify such content, enabling rapid circulation. This shift makes harm more widespread and longer-lasting than traditional systems can address. Hate speech reshapes interactions between social groups and challenges democratic institutions, social trust, and individual dignity over time. This report defines “invisible crime” as repeated hate speech that remains below legal thresholds but gradually produces measurable social and institutional harm. Invisible crime is used here as an analytical concept, not a formal criminal category.

Hate speech should not be understood merely as an illegal individual expression. It also operates as a process that gradually normalizes harm over time. The danger of hate speech lies not only in its content. It also lies in the chain reactions it can trigger. History and contemporary reality demonstrate that when specific groups are persistently portrayed as “threats,” “contaminants,” or “parasites,” such narratives gradually reshape public cognitive frameworks. Consequently, social exclusion and institutional discrimination may become more cognitively acceptable, potentially increasing the risk of broader social harm. In this regard, repeated hate narratives can function as early indicators of social instability. When dehumanizing language becomes normalized, public tolerance for exclusion and discrimination increases.

For policymakers, this means hate speech should be treated not only as an expression issue but also as a risk factor for democratic erosion. The United Nations Secretary-General has noted in the UN Strategy and Plan of Action to Counter Hate Speech that such discourse typically intensifies in contexts of social division, economic inequality, and political polarization, and may predict conflict escalation.¹

Invisible crime has three core features: it remains below criminal prosecution thresholds, recurs frequently enough to normalize discriminatory narratives, and produces gradual social or institutional consequences over time.

Unlike traditional crimes, these behaviors do not manifest as isolated violent incidents. Instead, through normalized and repeated expressions, they gradually reshape societal boundaries of “acceptable harm,” blurring the line between speech and potential criminal acts.

¹ United Nations, *United Nations Strategy and Plan of Action on Hate Speech*, <https://www.un.org/en/genocide-prevention/hate-speech/strategy-plan-action>

For operational purposes, invisible crime can be identified through the following indicators:

- Recurrence threshold: repeated dehumanizing or exclusionary narratives within a defined time frame (e.g., sustained repetition over months within a digital community);
- Normalization markers: increasing public tolerance for discriminatory framing, measured through engagement patterns, discourse shifts, or polling trends;
- Escalation trajectory: observable progression from coded language to explicit dehumanization;
- Institutional inaction: documented recurrence without legal, regulatory, or platform response.

Invisible crime transitions into prosecutable or visible crime when speech satisfies established legal criteria such as incitement to discrimination, hostility, or violence under applicable human rights law frameworks.

However, if hate speech indeed possesses this potential to evolve into “invisible crime,” then the critical question is: can existing research adequately explain how expressive behavior transforms into structural harm?

Existing research explains hate speech primarily through communication environments, group dynamics, group interaction patterns, and community structures. Related studies generally agree that the communication environment and group dynamics are key conditions driving the reinforcement and spread of hate speech. For example, some research points out that the echo chamber effect, formed in highly interactive communities, intensifies cognitive homogeneity among members, causing hate speech to be repeated, amplified, and gradually escalated within the group.²

However, this approach treats hate speech as a passive outcome of external conditions and overlooks its internal cultural mechanisms. The limitation of this theoretical orientation is its inadequate ability to fully explain why hate speech can be continuously replicated across different social groups and achieve long-term diffusion across national borders and social contexts. In other words, an environmentalist approach alone cannot account for why certain hate speech exhibits remarkable resilience and reproducibility across contexts.

While existing research emphasizes communication structures and interactive environments, it often centers its analysis on external conditions. This assumes that hate speech is mainly shaped by context, while overlooking its internal cultural mechanisms. This underlying premise overlooks the crucial fact that many instances of hate speech are not merely situational emotional outbursts, but rather discursive forms embedded within specific cultural structures.

² Vasu Goel et al., “Hatemongers Ride on Echo Chambers to Escalate Hate Speech Diffusion.”

PNAS Nexus, 8 February 2023, <https://doi.org/10.1093/pnasnexus/pgad041>

This report instead proposes an analytical framework centered on cultural mechanisms. This framework suggests that hate speech achieves high diffusivity due to its reliance on simplified, emotional expressions that significantly lower cognitive engagement thresholds. Its narrative structure relies on simplified “us-versus-them” framing and moralized labeling. Through repetition, it increases tolerance for extreme narratives and reduces exposure to dissenting perspectives.

Over time, this reinforcement strengthens group conformity and normalizes exclusionary beliefs. Under this mechanism, hate narratives form stable internal cycles that reinforce exclusionary and radicalized logic. Consequently, hate speech transcends mere expression to become an underlying mechanism with cultural self-replicating capacity.

Building on this theoretical foundation, this report analyzes how hate speech evolves from isolated expressions into normalized practices that blur the boundaries between speech and potential crime at the cultural systemic level. By placing the cultural characteristics of hate speech itself at the core of analysis, the study aims to reveal its cross-contextual propagation logic, long-term evolutionary pathways, and the resulting institutional risks and governance challenges.

Finally, it must be emphasized that the issue of hate speech is not merely a simple “speech” problem at the normative level. It involves the systemic tension between freedom of expression and anti-discrimination protections, the balance between positive and negative obligations under international human rights law, and the responsibility realignment stemming from the evolving role of digital platforms in the public sphere. This report adopts an interdisciplinary and exploratory approach, combining literature review and comparative policy analysis. It does not aim to provide exhaustive coverage of all relevant disciplines, but rather to develop a conceptual framework linking discourse, digital structures, and institutional outcomes.

Taken together, this section establishes “invisible crime” as a progressive and pre-legal form of harm, highlighting that harm can emerge before it becomes legally visible.

2. Hate Speech within the Framework of International Human Rights Law

Under international human rights law, hate speech is not automatically prohibited, as the legal framework balances freedom of expression with the prevention of harm. However, this balancing model presupposes that harm is discrete, identifiable, and legally attributable, which makes it difficult to address harms that emerge through repeated speech but accumulate into broader social or institutional consequences. From the perspective of “invisible crime,” this assumption is limited because it overlooks potential cumulative harm. Therefore, the central policy question is when expression becomes harmful enough to justify restriction.

This section examines how this framework addresses hate speech and why it struggles to capture forms of harm—such as invisible crime—that develop through repeated speech and produce cumulative consequences over time.

First, freedom of expression holds a foundational status in international law. Article 19 of the International Covenant on Civil and Political Rights (ICCPR) further confirms this right in a legally binding form.³

However, the ICCPR does not treat freedom of expression as an absolute right. Article 19 explicitly permits restriction of this freedom when necessary to protect the rights of others, national security, public order, public health, or public morals, provided such restrictions are prescribed by law.⁴ More critically, Article 20 requires states to “prohibit any advocacy of

³ OHCHR, *General Comment No. 34: Article 19 – Freedoms of Opinion and Expression*, 12 September 2011, <https://www.ohchr.org/sites/default/files/english/bodies/hrc/docs/GC34.pdf>

⁴ OHCHR, *International Covenant on Civil and Political Rights*, 16 December 1966, <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence.”⁵

This creates a dual obligation: states must avoid excessive censorship while also intervening when speech constitutes incitement to discrimination, hostility, or violence. Restrictions must satisfy legality, necessity, and proportionality. This reflects the core principle that rights must not infringe on the rights and dignity of others.

Within this framework, the UN Human Rights Committee further clarified in General Comment No. 34 that any restriction on freedom of expression must simultaneously satisfy the principles of legality, legitimate purpose, necessity, and proportionality. This means states cannot impose excessive censorship under the name of “combating hate speech,” nor can they allow speech that reaches the threshold of incitement to pose real risks to vulnerable groups.

Thus, from a normative perspective, “hate speech” does not equate to all offensive, insulting, or inflammatory expressions. Only when speech reaches the legal threshold of “inciting discrimination, hostility, or violence” and poses a real risk of harm in a specific context may it constitute speech subject to restriction or even prohibition.

To further refine this threshold, the Office of the United Nations High Commissioner for Human Rights issued the Rabat Plan of Action, proposing a “six-factor test” for determining whether hate speech incites⁶:

- Context
- Speaker
- Intent
- Content and form
- Extent of dissemination
- Likelihood and imminence

The Rabat Plan of Action requires a contextual assessment, evaluating speech based on context, intent, reach, and likelihood of harm.

By incorporating elements such as “intent” and “real danger” into its analytical framework, the Rabat Plan of Action provides a structured assessment tool for national legislation and judicial discretion. It also establishes relatively objective standards for civil society organizations engaged in monitoring, advocacy, and policy evaluation. However, when assessed through the framework

⁵ OHCHR, *International Covenant on Civil and Political Rights*, 16 December 1966, <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

⁶ OHCHR, *OHCHR and freedom of expression vs incitement to hatred: the Rabat Plan of Action*, <https://www.ohchr.org/en/freedom-of-expression>

of “invisible crime,” the limitation becomes apparent: it primarily addresses discrete acts of incitement, while failing to capture how repeated low-intensity speech gradually normalizes exclusion over time and reshapes the boundaries of acceptable discourse.

Within international human rights law, it is important to distinguish between binding legal instruments and interpretative or soft-law guidance. The International Covenant on Civil and Political Rights (ICCPR) constitutes binding international law for ratifying states, particularly Articles 19 and 20. In contrast, instruments such as the Rabat Plan of Action and General Comments issued by the Human Rights Committee function as non-binding interpretative frameworks, guiding implementation but not creating direct legal obligations.

Overall, international human rights law does not adopt a one-dimensional stance between freedom of expression and combating hate speech. Instead, it constructs a prudent balancing mechanism centered on the principle of proportionality and grounded in specific contexts. This mechanism safeguards the openness of public discourse while providing clear legal boundaries to prevent discriminatory incitement. However, this legal balancing framework becomes increasingly strained in digital environments, where incremental and algorithmically amplified harms challenge its underlying assumptions.

This section shows that international human rights law primarily focuses on discrete acts of incitement, which makes it difficult to address incremental and low-intensity harms such as invisible crime. One possible alternative is the adoption of risk-based monitoring approaches rather than expanding criminal thresholds. The aim is, therefore, not to expand criminal liability, but to develop risk-based monitoring tools that identify escalation while preserving freedom of expression. Such approaches focus on identifying patterns of repetition and escalation over time, allowing earlier intervention without requiring formal criminalization. Having established the legal limits, the next section examines how digital environments shape hate speech.

3. Structural Changes in Hate Speech in the Digital Age

The following analysis focuses on how digital platforms create conditions that facilitate the emergence and expansion of invisible crime through repetition, amplification, and persistence.

In the context of digital transformation, social media platforms have become the core spaces for contemporary public discourse. Digital platforms are not neutral communication infrastructures. They allow repeated hate speech to accumulate over time. Existing research indicates that platform engagement-optimized systems may indirectly amplify polarizing content that is emotionally charged, controversial, and extreme, as such content tends to generate higher engagement and longer interaction time. Under this mechanism, even non-mainstream hateful expressions can gain disproportionate visibility due to high engagement.

This “algorithmic amplification effect” structurally alters the dissemination logic of hate speech. First, in scale, its dissemination is no longer confined to specific offline communities but can rapidly spread across regions and nations. Second, in speed, the propagation cycle is significantly shortened, creating instantaneous diffusion and emotional resonance. Third, in persistence, the replicability and redistributability of digital content enable hate narratives to be continuously reproduced and reactivated across different platforms and contexts. Thus, this dynamic is intensified in digital environments through algorithmic amplification and cross-platform redistribution.

Within this dissemination structure, hate speech is not randomly distributed. Research indicates its primary victims are concentrated among groups already marginalized or disadvantaged within societal structures, including ethnic and religious minorities.⁷ The Council of Europe's manual for combating hate speech online notes that online hate speech often intertwines with existing social

⁷ United Nations Press, *Secretary-General Statement*, <https://press.un.org/en/2025/sgsm22674.doc.htm>

inequalities, reinforcing social exclusion through symbolic stigmatization and stereotypical narratives, and potentially translating into institutional discrimination.⁸

Thus, hate speech is not merely an expression of individual prejudice. It reflects and reinforces existing social inequalities, especially when amplified by digital platforms. Its dissemination mechanisms interact with social structures: existing inequalities provide fertile ground for hate narratives, while algorithmic systems offer channels for their amplification.

Within this context, reliance solely on national legal frameworks is insufficient to adequately address the issue. Beyond international and domestic legal norms, civil society organizations (CSOs) play a crucial complementary and catalytic role in combating hate speech governance. Their functions can be broadly categorized into three levels:

(I) Monitoring and Data Collection

First, CSOs systematically monitor hate speech across social media and public spaces, establishing databases and publishing annual or thematic reports. For instance, the Organization for Security and Co-operation in Europe (OSCE) operates dedicated monitoring mechanisms for hate crimes and hate speech, collecting data from member states and conducting trend analyses.⁹ Systematic data collection not only helps identify high-risk groups and areas but also provides an evidence base for policy formulation and judicial practice, thereby enhancing the precision and transparency of governance. This is especially relevant to invisible crime because repeated low-level harm requires long-term documentation rather than single-incident reporting.

(II) Education and Awareness-Raising

Second, education is regarded as a key pathway for long-term prevention of hate speech. The Council of Europe's manual *A Manual for Combating Hate Speech Online through Human Rights Education* emphasizes that fostering critical thinking skills and human rights awareness among youth can reduce societal acceptance of hate narratives at their source.¹⁰ Civil society organizations advance societal shifts in understanding the principles of diversity, equality, and respect through workshops, public discussions, campus initiatives, and media training. This erodes the cultural foundations of hate speech. This is particularly important for invisible crime, as it targets the normalization process through which repeated narratives become socially accepted.

⁸ Council of Europe, *Bookmarks: A Manual for Combating Hate Speech Online through Human Rights Education*, 2014, <https://book.coe.int/en/human-rights-and-democracy/6023-pdf-bookmarks-a-manual-for-combating-hate-speech-online-through-human-rights-education.html>

⁹ OSCE, *Hate Crime Reporting Platform*, <https://hatecrime.osce.org>

¹⁰ Council of Europe, *Bookmarks: A Manual for Combating Hate Speech Online through Human Rights Education*, 2014, <https://book.coe.int/en/human-rights-and-democracy/6023-pdf-bookmarks-a-manual-for-combating-hate-speech-online-through-human-rights-education.html>

(III) Policy Advocacy and Legal Reform

Third, at the policy level, civil society organizations advance more robust, human rights-compliant legal frameworks by submitting policy recommendations, participating in legislative hearings, and publishing evaluation reports. For instance, Human Rights Watch consistently publishes research reports on democratic risks, and the boundaries of free speech, exerting public oversight pressure on national legislative and law enforcement practices.¹¹ The core of such advocacy lies in finding an institutionalized balance between safeguarding freedom of expression and preventing incitement to violence, avoiding governance measures that either impose excessive restrictions or create regulatory vacuums. This is crucial for invisible crime, where harm emerges gradually and requires policy responses that go beyond single-event regulation.

In brief, hate speech in the digital age has evolved from an issue of individual expression into a systemic problem shaped by algorithmic mechanisms, social structures, and political contexts. Within this complex framework, legal norms, platform responsibilities, and civil society actions form a multi-layered governance structure. Only by integrating diverse governance tools based on system-level analysis can we effectively address the evolving challenges of hate speech in the digital sphere, particularly its transformation into “invisible crime” through repeated, normalized, and structurally embedded patterns of dissemination.

Building on these systemic conditions, digital platforms therefore create organizational conditions that allow invisible crime to develop through repeated exposure, amplification, and normalization across contexts. Based on this, the next section examines the underlying cultural and cognitive mechanisms that enable such dynamics to persist and intensify.

¹¹ Human Rights Watch, *Free Speech*, <https://www.hrw.org/topic/free-speech>

4 ● The Mechanisms Behind Hate Speech

Building on the previous section, this section explains how hate speech becomes persuasive and self-reinforcing at the cultural and cognitive level to accumulate into broader social harm over time.

Hate speech operates through cultural mechanisms. It labels groups and reduces complex social issues to emotionally charged moral judgments, mobilizing group identity and antagonism in ways that prioritize emotional resonance over factual accuracy.¹²

This simplified emotional expression increases conformity and makes hate speech easier to spread by reducing the need for critical evaluation and encouraging intuitive responses. On one hand, it lowers the threshold for critical thinking among audiences, making individuals more prone to accept and disseminate related narratives without verifying information.^{13 14}

On the other hand, this highly emotional form of expression renders hate speech itself a “manipulable resource,” making it easily adaptable and reusable by political actors, media, or platform environments for social mobilization, public opinion steering, or emotional manipulation.¹⁵ As a result, hate speech shifts from spontaneous expression to a reproducible communicative mechanism embedded within broader systems.

¹² George Lakoff, “What is Hate Speech?”, 14 September 2017, <https://george-lakoff.com/2017/09/14/what-is-hate-speech>

¹³ Martin Bruder and Laura Kunert, “The conspiracy hoax? Testing key hypotheses about the correlates of generic beliefs in conspiracy theories during the COVID-19 pandemic”, February 2022, <https://doi.org/10.1002/ijop.12769>

¹⁴ Scan Project, *Hate Speech Trends during the COVID-19 Pandemic in a Digital and Globalised Age*, <http://scan-project.eu/wp-content/uploads/sCAN-Analytical-Paper-Hate-speech-trends-during-the-Covid-19-pandemic-in-a-digital-and-globalised-age.pdf>

¹⁵ Radu Meza et al., “Targets of Online Hate Speech in Context. A Comparative Digital Social Science Analysis of Comments on Public Facebook Pages from Romania and Hungary,” April 2018, <https://www.cceol.com/search/article-detail?id=908843>

At the social group level, this mechanism of uncritical conformity manifests as selective attraction toward specific groups, reinforcing internal narrative cycles that are central to the formation of “invisible crime.”

Hate speech tends to attract individuals with lower exposure to pluralistic information environments or limited access to counter-narratives who lack a full understanding of the risks inherent in extremist ideologies.¹⁶ Such individuals are more likely to assimilate within group interactions and gradually internalize extreme narratives while gaining a sense of belonging and security.

As group membership expands, internal agreement increases and tolerance for extreme speech rises. This process forms a stable internal cycle, enabling hate speech to persist and develop into self-sustaining group dynamics.

The online and platform environment further reinforces and amplifies the above mechanisms. Social media's design logic, centered on instant interaction, emotional feedback, and high visibility, enables hate speech to trigger widespread engagement and emotional resonance within a short timeframe, thereby achieving spontaneous diffusion.¹⁷ ¹⁸ Under these embedded incentives, right-wing extremist users gradually form networked communities composed of loosely connected individuals.¹⁹ These communities share hateful narratives while often endorsing or glorifying violent tactics, drawing on the experiences and symbolic resources of authoritarian regimes. This creates a self-reinforcing pattern in which repeated exposure may increase the likelihood that online narratives will influence offline behavior.

Political and social events provide crucial external triggers for this cyclical mechanism. Take anti-immigrant hate speech as an example: populist narratives often exhibit highly adaptive “chameleon-like” characteristics. Within specific political contexts, exclusionary nationalist discourses can be repackaged as seemingly inclusive appeals for solidarity or security, thereby becoming incorporated into the policy platforms of various political actors. Following terrorist attacks or highly visible social incidents, the concentrated political and media discourse surrounding related issues inevitably provides a legitimizing framework for hate speech and its

¹⁶ Iztok Šori and Vasja Vehovar, “Reported User-Generated Online Hate Speech,” 19 August 2022, <https://doi.org/10.3390/socsci11080375>

¹⁷ Alexander Brown, “What is so special about online (as compared to offline) hate speech?,” May 19, 2017, <https://doi.org/10.1177/1468796817709846>

¹⁸ Karin Wahl-Jorgensen, “An Emotional Turn in Journalism Studies?,” 12 December 2019, <https://doi.org/10.1080/21670811.2019.1697626>

¹⁹ Brian Ballsun-Stanton et al., “Online Right-Wing Extremism: New South Wales, Australia,” 27 April 2021, <https://doi.org/10.3390/proceedings2021077018>

associated behaviors.²⁰ This makes such rhetoric more easily perceived as an “understandable” or even “reasonable” response. This legitimization process not only reinforces group conformity but also facilitates temporal and spatial overlaps and resonances between hate speech and real-world hate crimes.

Finally, from the perspective of the content itself, the cultural mechanism of hate speech is also highly correlated with its form of expression. Such discourse typically employs highly simplified, emotional, and easily comprehensible linguistic structures, thereby significantly lowering participation thresholds and enabling a broader range of individuals to enter and engage in interactions.²¹ Over time, the gradual effect of this cultural mechanism enables hate speech to spread across different groups and contexts, exhibiting multi-layered, structured, and progressively worsening expansion characteristics, gradually becoming a latent risk factor for criminal acts.

Taken together, these cultural, cognitive, and group-based mechanisms provide the micro-level foundation through which hate speech evolves into “invisible crime,” enabling narratives to become socially embedded and self-reinforcing.

Building on this, these cultural and cognitive mechanisms explain how repeated hate speech becomes normalized and socially embedded, forming the micro-level foundation of invisible crime. Building on these mechanisms, the next section examines their structural effects.

²⁰ Matthew L Williams et al., “Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime,” January 2020, <https://doi.org/10.1093/bjc/azz049>

²¹ Iztok Šori and Vasja Vehovar, “Reported User-Generated Online Hate Speech,” 19 August 2022, <https://doi.org/10.3390/socsci11080375>

5. In-Depth Analysis: Structural Mechanisms of Hate Speech, Information Segregation, and “Invisible Crime”

While the previous section focused on how hate speech becomes self-reproducing at the cultural and cognitive level, this part analyzes how these mechanisms translate into structural and institutional consequences, while examining system-level and platform-level processes that amplify “invisible crime” and illustrating how progressive digital exposure produces broader societal and institutional effects.

Within this framework, hate speech can no longer be understood as a single act but rather as part of a broader process in which repeated expressions gradually accumulate and reshape social and institutional outcomes.

5.1. From Single Acts to Cumulative Harm

Traditional legal approaches typically treat hate speech as an isolated expression. The specific review mechanism operates as follows: an individual makes a statement, judicial authorities assess its incitement potential, and then determine whether it constitutes a violation of the law. This logic is structured around the principle of “single act—single liability.”

However, with the widespread proliferation of digital platforms, the mechanisms of harm generation have fundamentally shifted. The commercial logic of digital platforms is rooted in the “attention economy.” Extensive research indicates that emotionally intense content attracts greater engagement and dissemination, with anger, fear, and moral condemnation proving particularly contagious. The study on online harassment indicates that controversial content is more likely to

trigger engagement and amplification.²² Under this mechanism, hate speech gains heightened visibility due to its emotional intensity, while rational or moderate expressions are often suppressed. It is important to note that algorithms themselves do not actively favor hateful content; they merely prioritize pushing “high-engagement content” based on user interaction rates. However, this structural result objectively amplifies the spread of hateful discourse.

5.2. Digital Amplification and Information Isolation

Real-world harm increasingly results not from a single statement, but from repeated exposure and algorithmic amplification. Traditional legal systems designed around single-act liability struggle to address this cumulative process because their perspective fails to capture the compounding and systemic processes through which harm is generated over time.

A report on hate speech further indicates that digital spaces can form “information echo chambers.”²³ When users are persistently exposed to highly homogeneous narratives, group stereotypes are reinforced, dissenting voices fade, and extreme views gradually become normalized. This “gradual normalization” constitutes a critical stage in the formation of invisible crime. Within highly segmented information environments, hateful expression follows an escalating pattern:

- Initial stage: suggestive language;
- Developing stage: derogatory labels;
- Advanced stage: dehumanizing narratives;
- Final stage: justification of violence.

5.3. Structural Violence and Invisible Crime

This process is similar to Johan Galtung's concept of “structural violence.”²⁴ Harm is no longer directly inflicted by a single perpetrator but is embedded within institutional arrangements and communication structures, exerting sustained influence through normalized, routine operations. However, the two are not identical.

²² George Beknazar-Yuzbashev et al., “Toxic Content and User Engagement on Social Media,” 2025, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2025/toxic_content_and_user_engagement_on_social_media_evidence_from_a_field_experiment

²³ Vasu Goel et al., “Hatemonsters Ride on Echo Chambers to Escalate Hate Speech Diffusion,” PNAS Nexus, 8 February 2023, <https://doi.org/10.1093/pnasnexus/pgad041>

²⁴ Karsten Müller and Carlo Schwarz, “Fanning the Flames of Hate: Social Media and Hate Crime,” 2018, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2018/fanning_the_flames_of_hate_social_media_and_hate_crime

Structural violence refers to institutional arrangements that systematically produce inequality or harm. Invisible crime, by contrast, focuses on the communicative processes that normalize exclusion before such arrangements become formalized. While structural violence describes harm embedded in systems, invisible crime emphasizes how repeated discourse gradually reshapes the social acceptance of harm.

It accumulates gradually through repetition and reproduction. Within this structure, hate speech need not persuade society at large but only establish consensus within the inner circle. Invisible crime thus transforms into normalized behavior within the group.

Long-term, the United Nations Development Programme notes that declining social cohesion erodes institutional trust.²⁵ When specific groups are persistently portrayed as “threats,” linguistic exclusion gradually translates into institutional consequences:

- Discriminatory policies gain public support
- Exclusionary laws become justified
- Democratic principles are eroded

5.4. European Illustrations

In the 2015–2016 refugee influx in Germany, research shows that in municipalities with higher Facebook usage, right-wing anti-refugee sentiment on social media predicted higher rates of violent attacks against refugees.²⁶ This pattern shows that harm does not result from isolated statements but from a progressive process of exposure. Specifically, sustained exposure to hostile narratives contributes to shifting social perceptions of antagonistic attitudes, which in turn reduces social and psychological inhibitions against aggression and expands the perceived boundaries of acceptable behavior. As these inhibitions weaken, hostility becomes increasingly legitimized, potentially increasing the likelihood that individuals translate online sentiment into offline action.

This case suggests an association between repeated online hostility and offline violence. Sustained exposure to hostile discourse fosters an environment in which violence becomes progressively more conceivable, socially tolerated, and ultimately actionable, demonstrating repeated low-level hostile narratives function as early-stage normalization mechanisms within the invisible crime framework. In this sense, it is not isolated extreme statements, but the sustained circulation of normalized hostility that constructs the conditions for violence. Such a dynamic directly reflects the core logic of “invisible crime,” whereby harm emerges not from discrete, easily identifiable acts, but

²⁵ United Nations, *TRUST IN A CHANGING WORLD: SOCIAL COHESION AND THE SOCIAL CONTRACT IN UNCERTAIN TIMES*, December 2024, https://social.desa.un.org/sites/default/files/inline-files/World%20Social%20Report_Dec2024.pdf

²⁶ Karsten Müller and Carlo Schwarz, “Fanning the Flames of Hate: Social Media and Hate Crime,” 2018, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2018/fanning_the_flames_of_hate_social_media_and_hate_crime

from diffuse, incremental processes that reshape behavioral norms and enable latent aggression to manifest in overt forms.

Another European case is the online anti-Muslim narratives in Italy, particularly observed in Facebook and Twitter discourse during the 2015–2018 migration period, where studies show that Muslim identity was frequently framed as a security threat in political and media-linked narratives.²⁷ Anti-Muslim discourse online often frames Muslim identity as an external threat and reinforces pre-existing stereotypes. These narratives do not necessarily incite specific violent acts but operate through gradual normalization processes. This reflects another dimension of invisible crime, where harm emerges through the gradual shifts in acceptable discourse rather than direct incitement. Over time, this sustained exposure reshapes public perception and social boundaries, making discriminatory positions appear increasingly reasonable, while “invisible crime” emerges through long-term discursive practices rather than explicit incitement.

Finally, the cyclical amplification of hate speech within digital structures not only constitutes the generative mechanism of invisible crime but also propels its progression from linguistic to institutional levels. Harm no longer manifests as isolated eruptions but accumulates persistently within institutional frameworks and communication logics, ultimately exerting profound impacts on democratic order and social integration.

While the above analysis focuses on European cases, these examples are used as illustrative contexts to reflect broader global dynamics, as similar patterns of digitally mediated hate speech and cumulative normalization can be observed across different regions and political systems.

5.5. Multi-Level Effects of Invisible Crime

The structural generation of invisible crime reflects the interaction of cognitive, social, and technological dynamics: prolonged exposure to homogeneous and emotionally charged narratives reinforces bias, group-level conformity amplifies internal consensus while marginalizing dissent, and platform algorithms accelerate visibility through engagement-based ranking. At the same time, legal limitations focusing on explicit incitement allow these cumulative processes to persist, enabling gradual normalization of exclusion and its translation into institutional outcomes.

Meanwhile, institutional underreporting, distrust in law enforcement, and absent social intervention further conceal the harm, making it difficult for victimized groups to obtain protection. Over time, linguistic exclusion gradually transforms into institutionalized discrimination: exclusionary policies gain public endorsement, while laws and public norms are reinterpreted to legitimize exclusionary practices.

²⁷ Alessandra Vitullo, “Online Hate Speech and Intersectionality: Evidence From the Anti-Muslim Narratives in Italy,” 26 June 2025, <https://academic.oup.com/edited-volume/60551/chapter-abstract/523642681>

Thus, invisible crime shaped by hate speech manifests not merely as interpersonal harm, but as long-term risks embedded within social structures and technological ecosystems. It is characterized by high concealability, cumulative effects, resistance to timely legal regulation, and the potential to permeate institutional and policy levels. These overlapping factors show that legal rules or platform regulation alone are not enough. Instead, it requires cross-level, multi-strategy interventions—including psychological education, group governance, algorithmic transparency, and social institutional reform—to effectively address its deep-rooted underlying mechanisms and long-term societal impacts.

In brief, this section demonstrates how sustained exposure, platform dynamics, and institutional limitations interact to transform repeated hate speech into structural harm, thereby constituting invisible crime at the societal level.

5.6. Risks of Regulatory Overreach

At the same time, it is important to recognize that expanding the analytical scope of hate speech also raises concerns regarding regulatory overreach. Excessive regulation or overly broad interpretations of harm may risk undermining freedom of expression, enabling selective enforcement, or politicizing content moderation. These risks highlight the importance of maintaining clear legal thresholds, procedural safeguards, and transparency mechanisms. A balanced approach therefore requires addressing cumulative harm while avoiding disproportionate restrictions on legitimate expression.

6. Policy Recommendations

The analysis shows that hate speech in digital environments operates cumulatively. It often remains below legal thresholds while gradually normalizing exclusion. Addressing this pattern requires preventive and systemic responses rather than relying solely on criminal prosecution.

1. *Integrate Pattern-Based Risk Assessment*

States and regulatory bodies should move beyond single-post evaluation by developing integrated monitoring systems that track repeated patterns of hate speech over time. These systems should incorporate escalation logs, coded-language detection, and longitudinal analysis to capture recurrence, escalation, and contextual targeting in a structured and cumulative manner. This can be operationalized through the development of cumulative risk indicators, such as frequency thresholds within defined time windows, escalation markers from coded to explicit language, and concentration patterns targeting specific vulnerable groups. These indicators should be integrated into regulatory and platform-level assessment tools to identify when repeated low-level speech signals emerge from structural risk.

This does not require criminalizing individual expressions, but recognizing that repeated speech below legal thresholds can, over time, accumulate into broader social or institutional harm.

2. *Strengthen Platform Accountability*

Digital platforms should implement structured transparency mechanisms such as public dashboards, researcher API access, transparency reports, and algorithmic audit trails that make algorithmic amplification processes auditable and accountable. This includes disclosing how engagement-based ranking systems prioritize content, establishing internal escalation monitoring systems during high-risk periods such as elections or crises, and introducing cumulative harm assessment models that evaluate patterns of repeated exposure rather than isolated posts. Platforms should also develop internal risk dashboards that track the amplification of high-risk narratives over time and trigger early intervention measures when escalation thresholds are reached because algorithmic amplification can intensify the cumulative effects that characterize invisible crime.

3. Support Independent Monitoring Mechanisms

Governments, in collaboration with CSOs, academic researchers, and independent oversight bodies, should institutionalize multi-stakeholder monitoring frameworks by establishing formal partnerships between platforms, academic researchers, and civil society organizations. These partnerships should operate through shared data access protocols, standardized reporting formats, and periodic risk assessment reports that identify recurring dehumanizing narratives, track their cross-platform evolution, and detect transition points from normalization to incitement, thereby functioning as early-warning infrastructures for both policy intervention and public awareness. Independent monitoring is necessary because invisible crime often escapes detection by law alone, requiring early-warning infrastructures that combine multiple perspectives.

4. Safeguard Freedom of Expression

All interventions must remain proportionate and grounded in international human rights standards by incorporating procedural safeguards such as independent oversight mechanisms, clear legal thresholds for intervention, and transparency requirements for both state and platform actions. The objective is not to suppress controversial or dissenting speech, but to prevent the cumulative normalization and standardization of dehumanizing narratives through narrowly targeted, evidence-based, and reviewable measures, while ensuring that measures addressing invisible crime do not compromise fundamental freedoms.

7 ● Conclusion

Based on the preceding analysis, it becomes evident that hate speech is not merely an issue of expression or a technical dispute over legal boundaries, but part of an accumulated process through which repeated discourse reshapes social norms and enables institutional harm. This allows invisible crime to persist as repeated low-intensity harm that remains difficult to regulate under existing legal frameworks.

Furthermore, the diffusion of hate speech and its evolution into invisible crime is a systemic process driven by interacting cultural, social, and technological mechanisms. These dynamics explain its persistence, adaptability, and the limitations of governance approaches focused solely on individual behavior or isolated interventions.

The central contribution of this report lies in identifying invisible crime as an intermediate stage between expression and institutionalized harm, and demonstrating that hateful discourse is not an isolated linguistic phenomenon but a cultural transmission mechanism that continuously operates within social structures.

If governance of hate speech overlooks its structural and accrual nature, interventions will struggle to effectively curb its spread and its translation into real-world actions. Ultimately, the key governance challenge is that harm can become socially significant before it becomes legally visible.

References

Alessandra Vitullo, "Online Hate Speech and Intersectionality: Evidence From the Anti-Muslim Narratives in Italy," 26 June 2025, <https://academic.oup.com/edited-volume/60551/chapter-abstract/523642681>

Alexander Brown, "What is so special about online (as compared to offline) hate speech?", May 19, 2017, <https://doi.org/10.1177/1468796817709846>

Brian Ballsun-Stanton et al., "Online Right-Wing Extremism: New South Wales, Australia," 27 April 2021, <https://doi.org/10.3390/proceedings2021077018>

Council of Europe, *Bookmarks: A Manual for Combating Hate Speech Online through Human Rights Education*, 2014, <https://book.coe.int/en/human-rights-and-democracy/6023-pdf-bookmarks-a-manual-for-combating-hate-speech-online-through-human-rights-education.html>

George Beknazar-Yuzbashev et al., "Toxic Content and User Engagement on Social Media," 2025, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2025/toxic_content_and_user_engagement_on_social_media_evidence_from_a_field_experiment

George Lakoff, "What is Hate Speech?", 14 September 2017, <https://george-lakoff.com/2017/09/14/what-is-hate-speech>

Human Rights Watch, *Free Speech*, <https://www.hrw.org/topic/free-speech>

Iztok Šori and Vasja Vehovar, "Reported User-Generated Online Hate Speech," 19 August 2022, <https://doi.org/10.3390/socsci11080375>

Johan Galtung and Tord Höivik, "Structural and Direct Violence: A Note on Operationalization," 1971, <https://journals.sagepub.com/doi/10.1177/002234337100800108>

Karsten Müller and Carlo Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime," 2018, https://warwick.ac.uk/fac/soc/economics/research/centres/cage/publications/workingpapers/2018/fanning_the_flames_of_hate_social_media_and_hate_crime

Karin Wahl-Jorgensen, "An Emotional Turn in Journalism Studies?", 12 December 2019, <https://doi.org/10.1080/21670811.2019.1697626>

Martin Bruder and Laura Kunert, "The conspiracy hoax? Testing key hypotheses about the correlates of generic beliefs in conspiracy theories during the COVID-19 pandemic", February 2022, <https://doi.org/10.1002/ijop.12769>

Matthew L Williams et al., "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime," January 2020, <https://doi.org/10.1093/bjc/azz049>

OHCHR, *General Comment No. 34: Article 19 – Freedoms of Opinion and Expression*, 12 September 2011, <https://www.ohchr.org/sites/default/files/english/bodies/hrc/docs/GC34.pdf>

OHCHR, *International Covenant on Civil and Political Rights*, 16 December 1966, <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

OHCHR, *OHCHR and freedom of expression vs incitement to hatred: the Rabat Plan of Action*, <https://www.ohchr.org/en/freedom-of-expression>

OSCE, *Hate Crime Reporting Platform*, <https://hatecrime.osce.org>

Radu Meza et al., "Targets of Online Hate Speech in Context. A Comparative Digital Social Science Analysis of Comments on Public Facebook Pages from Romania and Hungary," April 2018, <https://www.ceeol.com/search/article-detail?id=908843>

Scan Project, *Hate Speech Trends during the COVID-19 Pandemic in a Digital and Globalised Age*, <http://scan-project.eu/wp-content/uploads/sCAN-Analytical-Paper-Hate-speech-trends-during-the-Covid-19-pandemic-in-a-digital-and-globalised-age.pdf>

United Nations, *United Nations Strategy and Plan of Action on Hate Speech*, <https://www.un.org/en/genocide-prevention/hate-speech/strategy-plan-action>

United Nations, *TRUST IN A CHANGING WORLD: SOCIAL COHESION AND THE SOCIAL CONTRACT IN UNCERTAIN TIMES*, December 2024, https://social.desa.un.org/sites/default/files/inline-files/World%20Social%20Report_Dec2024.pdf

United Nations Press, Secretary-General Statement, <https://press.un.org/en/2025/sgsm22674.doc.htm>

Vasu Goel et al., "Hatemonsters Ride on Echo Chambers to Escalate Hate Speech Diffusion," PNAS Nexus, 8 February 2023, <https://doi.org/10.1093/pnasnexus/pgad041>



Solidarity With OTHERS

Belgium, 2026

www.solidaritywithothers.com

info@solidaritywithothers.com