Europe at a Crossroads: Governance, Business, and Policy Transformation in a Changing Global Economy. Proceedings of the International Conference EU-PAIR 2026

DOI:10.47743/eu-pair.2026-1-14

FREEDOM OF EXPRESSION, HATE SPEECH AND DIGITAL GOVERNANCE: CONTEMPORARY LEGAL CHALLENGES IN DEMOCRATIC SOCIETIES

ARMINE KHARATYAN
Armenian State University of Economics
Yerevan, Republic of Armenia
kharatyan.armine@asue.am

GEGHETSIK GRIGORYAN
Armenian State University of Economics
Yerevan, Republic of Armenia
ggrigorian08@gmail.com

VAHE TOROSYAN
International University of Applied Sciences in Armenia
Yerevan, Republic of Armenia
vaxinak91@mail.ru

Abstract
The rapid development of digital technologies and communication platforms has fundamentally transformed the exercise of freedom of expression and the dissemination of information in contemporary societies. At the same time, the growing prevalence of hate speech in online environments has generated significant legal challenges concerning the protection of human dignity, equality, non-discrimination, and democratic values. These developments require legal systems to identify appropriate mechanisms for addressing harmful forms of expression while preserving the fundamental right to freedom of expression.
This article examines the legal concept of hate speech and the criteria used to distinguish protected expression from speech that may legitimately attract legal responsibility. The research is based on doctrinal legal analysis, comparative examination of international and regional legal standards, and the study of contemporary judicial approaches. Particular attention is paid to the jurisprudence of international courts and the evolving regulatory frameworks governing digital platforms and online intermediaries.
The paper further analyses the increasing role of digital governance mechanisms, including content moderation policies, algorithmic amplification, and platform

accountability, in shaping modern approaches to harmful online content. It also explores recent developments in judicial practice, including the application of the six-factor threshold test of the United Nations Rabat Plan of Action as a structured methodology for assessing potentially harmful speech.

The study argues that effective regulation of hate speech requires a contextual and proportionate approach that simultaneously safeguards freedom of expression and protects individuals and groups from incitement to hatred, discrimination, or violence. It concludes by proposing legal and policy recommendations aimed at improving regulatory frameworks and strengthening accountability mechanisms in the digital environment.

Keywords: *freedom of expression; hate speech; digital governance; human rights; platform accountability.*

JEL Classification: K24; K33; K38.

1. INTRODUCTION

Freedom of expression is both an individual right and an institutional condition of democratic government. It enables the circulation of information and ideas, political pluralism, scrutiny of public power, and participation in matters of public concern. Its protection is therefore especially strong in relation to political debate, journalism, academic inquiry and other forms of public-interest expression. Yet freedom of expression is not absolute. It may come into conflict with human dignity, equality, privacy, personal security and the right to be protected against discrimination and violence. The central legal task is not to grant automatic priority to either side, but to identify criteria that distinguish protected, even offensive, speech from expression that may legitimately attract legal responsibility.

Digital communication has made that task substantially more complex. Social networks, video-sharing services, search engines and online news platforms allow a single user to reach a large and often indeterminate audience almost instantly. Visibility is shaped not only by the speaker's choices but also by recommender systems, ranking mechanisms, paid promotion, network effects and design features intended to maximise engagement. Content may be replicated, decontextualised and amplified long after its initial publication. Consequently, harmful expression in the digital environment cannot be understood solely as a bilateral relationship between a speaker and an addressee. Its social impact may be co-produced by the architecture and governance choices of online intermediaries.

The term 'hate speech' is widely used in legal and policy discourse, but it does not denote one universally defined offence. It covers expressions of different gravity, purposes and legal consequences. Some forms reach the high threshold of advocacy that constitutes incitement to discrimination, hostility or violence and may warrant criminal prohibition. Others may justify civil, administrative, professional or platform-based responses. Still others remain

legally protected, even when they are intolerant, disturbing or socially harmful, and are more appropriately addressed through counter-speech, education, media literacy and support for targeted communities (Council of Europe, 2022; Ghazaryan and Asatryan, 2021). Treating all these situations as a single category risks both under-protection and over-restriction.

This article addresses three interrelated questions. First, which legal criteria distinguish protected expression from hate speech that may legitimately be restricted, particularly where criminal liability is contemplated? Second, how do algorithmic distribution and platform moderation affect the assessment of reach, risk, and remedy? Third, what does the Armenian Court of Cassation's reliance on the six-factor Rabat threshold test in Case No. ED/1485/01/21 reveal about the domestic judicial application of international standards? The decision is examined as a focused case study rather than as the article's principal analytical framework. Its broader comparative relevance lies in demonstrating both the potential and the limitations of incorporating an international threshold test into domestic criminal adjudication.

The article's principal contribution is an integrated three-stage model of legal assessment. The first stage concerns contextual legal characterisation: the target, protected characteristic, communicative purpose, institutional setting and relationship between speaker and audience. The second stage concerns the threshold and form of response: criminal, civil, administrative, professional, platform-based or non-legal. The third one concerns the governance of dissemination: the role of algorithmic amplification, platform design, notice-and-action procedures, reason-giving and redress. By connecting substantive human-rights thresholds with the procedural architecture of the Digital Services Act, the model treats the Rabat factors not merely as possible grounds for restriction, but also as safeguards against mechanical or excessively broad interference with freedom of expression.

2. METHODOLOGY AND ANALYTICAL FRAMEWORK

2.1. Research Design

The study follows a qualitative legal research design combining doctrinal legal analysis, comparative legal analysis, systemic interpretation, normative mapping and a focused case-study method. These methods are used for different but connected purposes. Doctrinal analysis identifies the legal elements and safeguards governing restrictions on expression. Comparative analysis examines the interaction of United Nations, Council of Europe and European Union standards rather than undertaking a complete country-by-country comparison. Systemic interpretation connects substantive rules on incitement with procedural duties concerning content governance, transparency and redress.

The primary normative corpus comprises Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34 of the Human Rights Committee, the Rabat Plan of Action, Articles 10 and 17 of the European Convention on Human Rights (ECHR), the relevant case law of the European Court of Human Rights, Recommendation CM/Rec(2022)16 of the Committee of Ministers of the Council of Europe, and ECRI General Policy Recommendation No. 15. At European Union level, the analysis includes Council Framework Decision 2008/913/JHA, Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act, or DSA), and Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law supplies an additional lifecycle-based human-rights perspective.

Normative mapping is used to separate four questions that are frequently collapsed in debates on online hate speech: what substantive conduct is prohibited; which actor bears the relevant duty; which procedure governs the decision; and which remedy or safeguard is available. This distinction is particularly important because the DSA does not harmonise the substantive criminal law of hate speech. It regulates the governance of intermediary services while the illegality of particular content continues to depend on applicable Union or national law.

Table 1. Normative mapping of the principal legal sources

Source	Principal legal question	Primary addressee	Function in the analysis
ICCPR arts. 19–20 and Rabat Plan	When does advocacy become incitement, and when may criminal prohibition be justified?	States, legislators and courts	High substantive threshold; legality, intent, context and likelihood safeguards
ECHR arts. 10 and 17; ECtHR case law	When is interference prescribed by law, necessary and proportionate, and when do positive obligations arise?	States and domestic courts	Contextual proportionality review and protection of democratic debate

Source	Principal legal question	Primary addressee	Function in the analysis
CM/Rec(2022)16 and ECRI GPR No. 15	How should legal and non-legal responses be differentiated according to severity?	States, public officials, media, intermediaries and civil society	Graded response, prevention, remedies, counter-speech and stakeholder duties
EU Framework Decision 2008/913/JHA	Which serious racist and xenophobic expressions must be criminalised within its defined scope?	EU Member States	Minimum criminal-law approximation for public incitement and specified denial offences
Digital Services Act	How must intermediary services govern illegal content, platform rules, recommender systems and systemic risks?	Intermediary services, online platforms, VLOPs/VLOSEs and regulators	Procedural accountability, transparency, redress, risk assessment and scrutiny
AI Act and CoE AI Framework Convention	Which safeguards should govern AI-supported detection and moderation?	Public authorities and relevant AI actors	Risk-sensitive transparency, oversight, accountability, equality and impact management
Armenian Case No. ED/1485/01/21	How can the Rabat factors be applied in domestic criminal adjudication?	National courts	Focused case study of structured contextual reasoning and its limits

Source: authors' elaboration

2.2. Case-Study Protocol and Limitations

The Armenian component is based on Article 42 of the Constitution of the Republic of Armenia, Articles 329 and 330 of the Criminal Code, and the decision of the Criminal Chamber of the Court of Cassation in Case No. ED/1485/01/21, delivered on 23 April 2026. Article 329 addresses public speech aimed at inciting or promoting hatred, discrimination, intolerance or enmity on

the grounds enumerated by the Code, whereas Article 330 addresses public calls for violence, the public justification or advocacy of such violence, and the dissemination of related materials on comparable grounds, subject to the statutory conditions. The case itself concerned the aggravated form of the offence under Article 330(2)(2). This distinction is central to the analysis: domestic law separates hate-based inciting speech from the more specific communicative act of calling for violence, even though the two categories may overlap in particular cases. The case analysis separates three levels: the facts and procedural history established by the judgment; the Court's own legal propositions; and the authors' assessment of their broader significance. The Rabat factors operate here as an analytical framework within the case-study method, not as an independent empirical research method.

The article does not compare the complete legislation of multiple jurisdictions, quantify the prevalence of online hate speech, or test the performance of a particular commercial moderation system. Nor does it treat a single domestic judgment as evidence of a settled regional trend. Its objective is normative and explanatory: to formulate a coherent path from legal characterisation to a proportionate remedy while identifying the additional duties and safeguards created by digital dissemination.

3. THE LEGAL CONCEPT OF HATE SPEECH AND THE NEED FOR A GRADED RESPONSE

3.1. A Broad Concept, not a Single Offence

The absence of a single binding international definition is not merely terminological. Legal systems differ in the protected characteristics they recognise, the conduct they prohibit, the mental element they require, and the degree of likely harm that must be shown. Recommendation CM/Rec(2022)16 adopts a broad human-rights-based understanding that encompasses forms of expression which incite, promote, spread or justify violence, hatred or discrimination against a person or group, or denigrate them, on the basis of real or attributed personal characteristics or status. The Recommendation simultaneously insists on differentiated responses according to the severity, context and likely impact of the expression (Council of Europe, 2022).

This approach is preferable to treating hate speech as a self-executing legal label. A statement may fall within a social or policy understanding of hate speech without satisfying the elements of a criminal offence. Conversely, a threat or public call for violence may be criminally punishable even if it is not directed against a group defined by a protected characteristic. Legal characterisation must therefore proceed from the applicable offence or cause of action, not from the prior assumption that all harmful expressions belong to one legal category.

The normative debate also cautions against simple answers. Brown (2015) shows that 'hate speech law' is not one homogeneous regulatory object, but a cluster of distinct legal restrictions that require different justificatory arguments. Waldron (2012) emphasises the harm that public denigration causes to the social standing and assurance of members of vulnerable groups. Heinze (2013), by contrast, warns that broad prohibitions in the sphere of democratic public discourse may suppress participation and viewpoint diversity. Lepoutre (2020) further questions whether the expressive value of prohibition, standing alone, can justify bans where counter-speech may communicate public condemnation. Armenian legal scholarship likewise warns that broad or indeterminate hate-speech rules may be applied disproportionately and should therefore be tested through legality, necessity and proportionality (Odabashyan, 2020; Ghazaryan and Asatryan, 2021). These positions are not entirely incompatible. They support a distinction between direct incitement, targeted harassment and exclusionary conduct on the one hand, and general, even deeply objectionable, opinions in public debate on the other. The legal response should correspond to the communicative act and the risk it creates, rather than to the offensiveness of the viewpoint alone.

3.2. Distinguishing Adjacent Legal Categories

Hate speech must be distinguished from insult, defamation, threat and hate-motivated crime. Insult primarily concerns an attack on an individual's dignity; defamation generally concerns the dissemination of a false factual allegation capable of harming reputation. A threat communicates an intention to cause harm or reasonably creates fear, while a public call for violence seeks to influence others to engage in violent conduct. A hate-motivated crime involves an independently punishable act, such as assault or property damage, committed with a bias motive. In hate-speech cases, by contrast, the immediate object of legal assessment is the communication itself.

The distinctions may overlap in practice, but they remain legally significant. A statement targeting an individual public official may be insulting, threatening or an incitement to violence without constituting classic group-based hate speech. Criticism of a religion, ideology, political party or public institution must be distinguished from advocacy of discrimination or violence against persons associated with that belief or institution. Likewise, the use of harsh or humiliating language does not automatically prove an intention to incite. The decision-maker must identify the target, the relevant characteristic, the conduct encouraged, and the connection between the expression and the response sought from the audience.

This is particularly important in criminal law, where legal certainty and strict interpretation apply. Terms such as 'hatred', 'hostility', 'propaganda', 'promotion' and 'incitement' cannot function as interchangeable moral

assessments. Incitement requires a communicative relationship between the expression and discriminatory, hostile or violent conduct by the audience. The more serious the sanction, the more precise the proof required regarding intent, context and likely impact (UN Human Rights Committee, 2011; OHCHR, 2012).

3.3. Protected characteristics and legal certainty

International instruments traditionally refer to race, colour, national or ethnic origin and religion, while contemporary European standards recognise a broader and non-exhaustive range of characteristics, including sex, gender, gender identity, sexual orientation, language, disability, age and other personal status (Council of Europe, 2022; ECRI, 2015). An open list can respond to emerging and localised forms of discrimination, but it also requires principled interpretation. If every temporary professional, political or institutional affiliation is treated as equivalent to historically marginalised status, hate-speech provisions may be transformed into devices for shielding powerful organisations or public authorities from criticism.

The identification of a protected characteristic should therefore be accompanied by analysis of vulnerability, patterns of discrimination, power relations and the effect of the expression on equal participation in social life. Domestic legislation may legitimately protect additional grounds, including political opinion, but courts should explain the legal basis and the consequences of doing so. The mere fact that a statement concerns politics does not make it protected, just as the mere reference to a political identity does not automatically convert it into group-based hate speech.

Table 2. A graded model of responses to harmful expression

Level	Indicative legal character	Primary responses	Essential safeguards
I. High threshold	Intentional advocacy constituting incitement to discrimination, hostility or violence; direct and serious public calls for violence	Criminal prohibition as ultima ratio; urgent protective measures where strictly necessary	Accessible and foreseeable law; proof of intent; contextual analysis; likelihood assessment; proportional sanction; judicial review

Level	Indicative legal character	Primary responses	Essential safeguards
II. Intermediate	Unlawful or rights-infringing expression below the criminal threshold	Civil, administrative, professional or targeted platform measures; right of reply; compensation; removal or visibility reduction where justified	Necessity and proportionality; reason-giving; distinction between fact and opinion; effective appeal; protection of journalism, research and counter-speech
III. Lawful but harmful	Intolerant, prejudicial or degrading expression that remains legally protected	Counter-speech, education, media and digital literacy, support for targeted persons, voluntary moderation and community standards	No coercive censorship; viewpoint neutrality; transparency; inclusion of affected communities

Source: authors' elaboration

4. INTERNATIONAL AND EUROPEAN STANDARDS FOR RESTRICTING HATE SPEECH

4.1. The Combined Operation of Articles 19 and 20 ICCPR

Article 19 ICCPR protects the freedom to seek, receive and impart information and ideas of all kinds. Restrictions are permissible only when provided by accessible and foreseeable law, directed to a legitimate aim recognised by the Covenant, and necessary and proportionate to that aim. The requirement of legality is qualitative: a formal statutory basis is insufficient if the rule is too vague to guide conduct or confers uncontrolled discretion on enforcement authorities (UN Human Rights Committee, 2011, paras. 22–36).

Article 20(2) requires States to prohibit by law advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. The provision does not require the prohibition of every intolerant or offensive opinion. Its elements include advocacy of hatred, an inciting relationship to audience conduct, and a sufficient risk of discrimination, hostility or violence. Articles 19 and 20 must be read together. Even where prohibition is mandatory, the content and application of the law must remain compatible with legal certainty, necessity and proportionality.

This combined reading prevents two opposite errors. The first is to treat freedom of expression as excluding all positive duties to protect targeted persons and groups. The second is to treat Article 20 as an autonomous license for broad criminalisation. The Rabat Plan of Action was developed precisely to reduce these risks by setting a high threshold and requiring structured analysis of the circumstances in which advocacy becomes incitement (OHCHR, 2012, para. 29).

Within the European Union, Council Framework Decision 2008/913/JHA requires criminalisation of specified forms of public incitement to violence or hatred directed against groups or members of groups defined by race, colour, religion, descent, or national or ethnic origin. Its narrower criminal-law scope should not be confused with broader policy definitions of hate speech. The Framework Decision confirms that criminal liability must remain tied to precisely defined conduct and protected grounds, while national systems may employ other proportionate measures for harmful expression falling below that threshold.

4.2. Articles 10 and 17 ECHR and the ECtHR's Contextual Approach

Under Article 10 ECHR, freedom of expression applies not only to information and ideas that are favourably received, but also to those that offend, shock or disturb. This protection reflects the demands of pluralism, tolerance and broadmindedness in a democratic society (Handyside v. the United Kingdom, 1976, para. 49). Interference must be prescribed by law, pursue a legitimate aim and be necessary in a democratic society. The European Court of Human Rights examines the expression in context, including its contribution to public debate, the speaker's position, the target, form, medium, audience and likely consequences.

Article 17, which prohibits the abuse of rights, may exclude from Convention protection expression aimed at destroying the rights and values guaranteed by the Convention. Its use should remain exceptional and confined to extreme cases (Perinçek v. Switzerland [GC], 2015, para. 114). Borderline cases are better examined under Article 10(2), where the State must demonstrate a pressing social need and proportionality. Automatic reliance on the label 'hate speech' risks removing speech from meaningful judicial scrutiny.

The Court's case law also requires attention to communicative role. In Jersild v. Denmark (1994, paras. 31, 33-35), the Court distinguished racist statements made by interviewees from the journalist's critical presentation of those statements. The distinction is highly relevant online, where automated tools may fail to separate original advocacy from quotation, documentation, condemnation, satire or academic analysis. Context is not an optional mitigating factor; it is part of the legal characterisation of the expression.

The Convention framework includes positive obligations. In Beizaras and Levickas v. Lithuania (2020, paras. 125-129), the Court found that the authorities had not responded effectively to serious homophobic comments and calls for violence posted online. In Budinova and Chaprazov v. Bulgaria (2021, paras. 62-67) and Behar and Gutman v. Bulgaria (2021, paras. 67-70), the Court recognised that extreme group-directed stereotyping may reach the threshold of Article 8 by affecting the identity, dignity and social participation of group members. These judgments show that protection against hate speech is not reducible to punishment of speakers; it also requires accessible remedies, effective investigation and equality in enforcement.

The Court has approached intermediary and account-holder responsibility through fact-sensitive analysis. In Delfi AS v. Estonia (2015, paras. 142-162), liability concerned a professionally managed commercial news portal, manifestly unlawful comments, the platform's control and economic role, and the effectiveness of remedies. In Sanchez v. France (2023, paras. 161-192), the Grand Chamber considered the responsibility of a politician who failed to remove unlawful third-party comments from a publicly accessible social-media page used during an election campaign. Neither judgment establishes universal strict liability for all user-generated content. Both require attention to the actor's role, knowledge, control, context and the nature of the comments.

4.3. The Six-Factor Rabat Threshold Test

The Rabat Plan of Action proposes six interdependent factors for determining whether expression reaches the threshold of incitement: (1) the social and political context; (2) the speaker's position or status; (3) intent; (4) content and form; (5) the extent of dissemination; and (6) the likelihood, including imminence, of harm (OHCHR, 2012, para. 29). The test is designed for a high threshold, especially where criminal sanctions are contemplated. It therefore serves a dual function: identifying genuinely dangerous incitement and preventing the use of vague hate-speech provisions against protected expression.

No factor should be treated as conclusive in isolation. A powerful speaker's status does not replace proof of intent. Wide dissemination does not by itself demonstrate a real likelihood of harm. Violent or degrading vocabulary may be evidence of dangerousness, but it does not automatically establish incitement. The decision-maker must explain how the factors interact in the particular case. This is especially important in digital environments, where reach may be produced by platform recommendations rather than by the speaker alone, and where apparent content may be detached from its original context.

The test should also be connected to the choice of remedy. A finding that expression is harmful does not predetermine criminal liability. The degree of intent, reach and likelihood may instead support civil protection, administrative

enforcement, professional discipline, platform intervention or non-coercive responses. Recommendation CM/Rec(2022)16 reinforces this differentiated model and places criminal law at the highest end of a broader continuum of measures (Council of Europe, 2022).

Table 3. The Rabat factors and their digital-governance relevance

Rabat factor	Core legal inquiry	Additional digital consideration
Context	Political tension, discrimination patterns, vulnerability and surrounding events	Coordinated campaigns, platform-specific norms, virality and cross-platform migration
Speaker	Authority, influence, institutional role and relationship with the audience	Verified or high-reach accounts, paid promotion, network centrality and control over a public page
Intent	Purpose to incite discrimination, hostility or violence	Use of coded language, repeated targeting, coordination and behavioural signals; caution against inferring intent from automation alone
Content and form	Meaning, tone, directness, imagery, repetition and call to action	Memes, multimodal content, irony, quotation, synthetic media and context collapse
Extent	Public nature, scale, frequency, audience and means of dissemination	Recommendation systems, personalised ranking, resharing, microtargeting and actual–not merely theoretical–reach
Likelihood and imminence	Real probability that the audience will engage in prohibited conduct	Speed of diffusion, audience susceptibility, offline mobilisation signals, and whether mitigation reduced the risk

Source: authors' elaboration

5. DIGITAL GOVERNANCE AND PLATFORM ACCOUNTABILITY

5.1. Substantive Illegality, Platform Rules and Systemic Governance

The legal debate on online intermediaries must distinguish liability for a particular item of content from accountability for the systems through which content is governed and distributed. A platform is not legally responsible for every unlawful statement posted by a user merely because it hosts or recommends information. It may nevertheless be required to maintain transparent procedures, respond to sufficiently precise notices, give reasons for restrictions, provide redress, assess systemic risks and subject its governance arrangements to scrutiny. Platform accountability is therefore broader than content liability, but it is also legally distinct from the speaker's responsibility.

The DSA preserves conditional exemptions from liability for intermediary services and prohibits the imposition of a general monitoring or active fact-finding obligation (arts. 4–8). At the same time, it establishes differentiated due-diligence duties according to the function and scale of the service. This architecture seeks to avoid both blanket immunity and universal strict liability. The latter would create powerful incentives to remove lawful but controversial expression, particularly where the applicable legal rule is uncertain.

The distinction between illegal content and content that merely violates private terms of service is central. Under the DSA, whether content is illegal depends on applicable Union or national law; the Regulation does not create a general substantive definition of illegal hate speech. Platforms may adopt broader community standards, but users should be able to identify whether a restriction was based on law, contract, or a systemic-risk measure. Conflating these grounds obscures the applicable test, the responsible decision-maker and the available remedy.

5.2. Notice-and-Action, Reason-Giving and Redress

Article 14 DSA requires providers to explain significant restrictions in their terms and conditions and to apply those restrictions diligently, objectively and proportionately, with due regard to the rights and legitimate interests of all parties. Article 16 requires hosting services to establish accessible notice-and-action mechanisms for specific items alleged to be illegal. Notices must be sufficiently precise and substantiated to permit an informed assessment; their processing must be timely, diligent, non-arbitrary and objective. The procedure is therefore not a notice-and-automatic-removal model. It is a notice-and-action system in which the response depends on a legally and factually reasoned assessment.

Article 17 requires a statement of reasons when a hosting service removes, disables access to, demotes, restricts monetisation, suspends an account or otherwise limits content for reasons of illegality or incompatibility with its

terms. Articles 20 and 21 add internal complaint-handling and certified out-of-court dispute settlement. These safeguards are essential where private moderation decisions affect participation in political, journalistic, academic or human-rights discourse. A rights-compatible system must protect both persons exposed to illegal content and users whose lawful expression is restricted in error.

Transparency obligations also support accountability at an aggregate level. Articles 15 and 24 require reporting on content moderation, including the use of automated tools, while Article 17 decisions feed a public statements-of-reasons database. For very large online platforms and very large online search engines (VLOPs and VLOSEs), Articles 37, 40 and 42 add independent audit, data access and enhanced reporting. These mechanisms make it possible to examine recurring errors, inconsistent treatment across languages, and the relationship between moderation policy and platform design.

5.3. Recommender Systems and Systemic Risks

The DSA also addresses the architecture of visibility. Article 27 requires transparency concerning the main parameters of recommender systems. For VLOPs and VLOSEs, Articles 34 and 35 require assessment and mitigation of systemic risks associated with the design or functioning of the service, including risks to fundamental rights, civic discourse, electoral processes and public security. Article 38 requires an option for a recommender system not based on profiling. The regulatory focus thus extends beyond the binary question whether content remains online to the ways in which ranking, recommendation, advertising and repeated exposure shape harm.

Algorithmic amplification changes the evidential meaning of the Rabat factor concerning extent. Content that is formally public may have negligible exposure, while a post initially addressed to a limited audience may acquire extensive reach through recommendation, paid promotion, coordinated sharing or cross-platform migration. Where available, decisions should therefore consider actual dissemination: impressions, resharing patterns, audience composition, duration of availability, repetition, monetisation and the contribution of automated recommendation.

Amplification also complicates attribution. The speaker remains responsible for the content and intended dissemination; the platform may bear separate governance duties arising from control over distribution systems, knowledge of systemic risks and capacity to mitigate them. The platform's recommender system does not prove the speaker's intent, and the speaker's intent does not determine whether the platform complied with its procedural or systemic obligations. Keeping these legal bases separate is necessary for both fairness and effective regulation.

Table 4. DSA duties relevant to online hate-speech governance

DSA mechanism	Relevant provisions	Human-rights function	Implication for hate-speech governance
Liability architecture and no general monitoring	Arts. 4–8	Prevents automatic attribution of all user content to intermediaries and limits incentives for over-removal	Platform duties must be analysed separately from speaker liability
Terms, notice-and-action and reasons	Arts. 14, 16–17	Legal certainty, objective processing and intelligible justification	Distinguish illegality, terms-of-service violations and risk-mitigation measures
Appeal and out-of-court redress	Arts. 20–21	Effective remedy and correction of erroneous restrictions	Affected users and notifiers require accessible contestation procedures
Transparency and moderation data	Arts. 15, 24, 42	Public scrutiny of moderation practices and automation	Report outcomes by legal ground, content category, detection method and type of restriction
Recommender transparency and user choice	Arts. 27, 38	User autonomy and visibility of ranking logic	Assess how profiling and recommendation alter reach and exposure
Systemic-risk assessment and mitigation	Arts. 34–35	Protection of fundamental rights, civic discourse and public security	Examine design incentives, coordinated abuse, language disparities and proportional mitigation
Independent audit and data access	Arts. 37, 40	External verification and evidence-based oversight	Enable research into prevalence, amplification, error patterns and mitigation effectiveness

Source: authors' elaboration

5.4. Proportionality and a Ladder of Interventions

Rights-compatible governance does not require maximal removal. Where content is manifestly illegal or presents an urgent threat to life or safety, removal or disabling access may be necessary. In other cases, proportionate measures may include restricting paid promotion, reducing recommendation, limiting resharing, adding context, imposing temporary functionality limits, age-appropriate restrictions, or preserving content for evidential and research purposes while limiting its circulation. The selected measure should correspond to the legal basis and demonstrated risk.

This ladder of interventions must not become an informal system of invisible censorship. Demotion and de-amplification can significantly affect public participation even when content remains technically accessible. Such measures should therefore be covered by clear rules, reason-giving, transparency reporting and effective appeal. The same principle applies to co-regulatory arrangements: public authorities may set procedural and risk-management duties, but should not transfer vague censorship mandates to private actors or to infrastructure providers that lack meaningful control over the specific content.

6. AUTOMATED MODERATION AND ARTIFICIA INTELLIGENCE

6.1. The Limits of Automated Classification

Automated systems can assist in detecting large volumes of potentially harmful content, but legal qualification cannot be reduced to lexical matching or statistical similarity. Hate speech is context-dependent and may be expressed through ambiguity, coded language, irony, images, historical references or local political meanings. The same words may constitute incitement, quotation, condemnation, satire, journalism or academic analysis. Research also demonstrates substantial disagreement among legally trained annotators when they apply different national laws to the same material (Korre et al., 2024, p. 4).

Two complementary studies illustrate why legal concepts should not be collapsed into a single binary label. Khurana et al. (2022, pp. 176-191) propose modular criteria concerning target groups, dominance, speaker characteristics, forms of negative group reference and potential effects, allowing definitions to be adapted transparently to the purpose of a specific task. Zufall et al. (2022, pp. 53-64) show that decomposing legal assessment into sub-decisions concerning the target group and the targeting conduct can improve both annotation consistency and the transparency of automated outcomes when compared with an end-to-end classification of punishability. These approaches support structured legal reasoning, but do not eliminate the need for qualified human judgment.

Errors are not evenly distributed. Systems trained on dominant-language datasets may perform poorly in smaller languages and dialects. They may over-identify the speech of minority communities or fail to recognise coded attacks

directed against them. False positives restrict lawful expression; false negatives leave targeted persons exposed. The legal relevance of an error depends not only on accuracy rates in the abstract, but also on the affected right, the reversibility of the decision, the availability of human review and the distribution of errors across groups.

Automated tools are particularly unsuitable as sole decision-makers where the issue concerns political speech, journalistic reporting, human-rights documentation, satire or the criminal threshold of incitement. In such cases, the system may prioritise content for review, identify patterns of coordinated dissemination or assist with evidence preservation, but the final legal assessment should remain with a qualified human decision-maker.

6.2. Human Oversight, Explanation and Impact Assessment

Meaningful human oversight requires more than the formal presence of a reviewer. The reviewer must have access to the relevant context, authority to depart from the automated recommendation, sufficient time, appropriate language and cultural competence, and protection from performance incentives that merely reproduce the machine's output. High-impact moderation systems should be audited for false-positive and false-negative rates, disparate effects on protected groups, handling of appeals and performance in the languages in which they are deployed.

Reason-giving is equally important. A user should be informed whether content was restricted because it was considered illegal, because it violated contractual terms, or because a risk-mitigation measure was applied. The reason should identify the rule, the key facts and the available remedy without disclosing information that would enable abuse of detection systems. Aggregate transparency reports should explain the role of automation, the number and outcome of appeals, and the frequency with which human review reverses automated decisions.

The AI Act contributes a risk-based regulatory framework, but it should not be presented as a complete code for online content moderation. Moderation and recommendation systems are not automatically classified as high-risk, and requirements such as human oversight cannot be generalised beyond the categories to which the Regulation makes them applicable. Its broader concepts of risk management, data governance, transparency and AI literacy remain relevant where the legal conditions are met. Recent legal scholarship similarly treats AI-supported hate-speech moderation as a field requiring calibrated legal safeguards, not technical accuracy alone (Balavadze, 2025). The Council of Europe Framework Convention on Artificial Intelligence adds a lifecycle-based human-rights perspective through principles of transparency and oversight, accountability, equality and non-discrimination, and risk and impact management. In the specific field of online speech, however, the DSA and

general human-rights law remain the more direct sources of procedural protection.

7. DOMESTIC JUDICIAL APPLICATION OF THE RABAT TEST: THE ARMENIAN COURT OF CASSATION

7.1. Case No. ED/1485/01/21: Facts and Holdings

Case No. ED/1485/01/21 concerned statements made by a schoolteacher during a lesson in the presence of pupils. The statements included explicit calls for lethal violence against the Prime Minister of Armenia and members of his family. The first-instance court convicted the defendant under Article 330(2)(2) of the Criminal Code, concerning public calls for violence committed through the use of an official or functional position. The Criminal Court of Appeal quashed the conviction and acquitted the defendant. On 23 April 2026, the Criminal Chamber of the Court of Cassation quashed the appellate judgment and restored the legal effect of the first-instance judgment (Court of Cassation of Armenia, 2026).

The Court of Cassation addressed the relationship between political expression and criminally punishable calls for violence. It recognised the importance of broad protection for political debate and criticism, but held that political subject matter does not immunise expression whose communicative content and context amount to a call for violence. The Court also rejected an exclusively formal understanding of 'use of official position'. It considered the teacher's professional role, the institutional setting of the classroom and the authority exercised over pupils relevant to the aggravated form of the offence.

Most importantly for the present article, paragraph 13.1 of the judgment reproduced the six Rabat factors: context; the speaker's identity or status; intent; content and form; extent of dissemination; and likelihood, including imminence, of harm. The Court did not create a separate seventh factor of 'audience vulnerability'. Rather, the pupils' age and relationship to the teacher were relevant within the context, speaker-status and likelihood analysis. The judgment also stated that the offence of a public call for violence is formal in the sense that the occurrence of actual violence or another material consequence is not required for completion of the offence.

The Court's analysis nevertheless did not reduce the offence to the presence of prohibited words. It examined the institutional setting, the relationship between speaker and audience, the nature of the statements and their potential influence. This combination is important: the absence of a required material consequence does not dispense with the need to demonstrate that the communication is genuinely a public call for violence and that criminal interference with expression is justified.

7.2. Comparative Significance and Doctrinal Limits

The decision's principal contribution is methodological. It moves the assessment from isolated words towards structured contextual reasoning and expressly incorporates an international threshold test into domestic criminal adjudication. For other jurisdictions, the case illustrates how non-binding international guidance can become operational through judicial interpretation without being treated as a substitute for the statutory elements of the offence.

Its broader significance should, however, be stated with precision. The case concerned public calls for violence against identified persons in a classroom setting. It was not a case concerning the liability of a digital platform, recommender system or online intermediary. Nor does it conclusively establish that the targeted public official and family members constituted a protected group in the classic international-law sense of group-based hate speech. The domestic statutory framework referred to political or other views, but the factual target remained specific individuals. The decision is therefore best characterised as an authoritative ruling on public calls for violence and the limits of political expression, informed by international incitement criteria.

A second issue concerns the relationship between a formal offence and the Rabat factor of likelihood and imminence. Because a material result is not an element of the offence, the prosecution need not prove that violence actually occurred. Yet likelihood remains relevant to the expressive nature of the conduct, the necessity of criminal intervention and the proportionality of punishment. If likelihood is treated as irrelevant, the Rabat test becomes ornamental; if it is treated as a requirement of completed harm, the domestic offence is altered. The more coherent approach is to use likelihood as evidence that the expression had a genuine inciting or calling character and that criminal law, rather than a less restrictive response, was warranted.

A third limit is evidential. Domestic courts should not merely list the six factors. They should identify the facts supporting each factor and explain their interaction. Speaker authority, audience age, repetition, publicity and political tension may increase risk, but none replaces proof of intent or the legal elements of the offence. Properly applied, the Rabat framework can strengthen both protection from violence and protection against arbitrary criminalisation.

8. DISCUSSION: AN INTEGRATED RIGHTS-BASED GOVERNANCE FRAMEWORK

8.1. Stage One: Legal Characterisation Before Legal Sanction

The analysis supports a three-stage model. The first stage is legal characterisation: who or what is targeted, on which ground, with what communicative purpose, in what institutional and social setting, and through what relationship between speaker and audience? This inquiry must distinguish group-directed incitement from threats, defamation, insult, harassment, public

calls for violence and hate-motivated offences. The expression should not be assigned a sanction merely because it falls within a broad policy definition of hate speech.

A broad or 'umbrella' definition may be useful for policy coordination, education, monitoring and victim support, provided that it does not operate as a self-executing basis for punishment. Each form of criminal, administrative or civil liability must rest on a separate, precise and foreseeable legal rule. This separation protects legal certainty and reduces the risk that a general commitment to combat hate speech is used to suppress criticism of governments, public officials, religious beliefs or influential social actors.

8.2. Stage Two: Threshold, Remedy and Institutional Safeguards

The second stage asks whether the expression reaches criminal incitement, another form of unlawful expression, or the category of lawful but harmful speech. Criminal responsibility should remain an ultima ratio response. It may be justified where the offence is foreseeable, intent is established, the expression genuinely advocates or calls for prohibited conduct, and the contextual evidence supports a real likelihood of discrimination, hostility or violence. A court should identify each relevant Rabat factor and explain how it relates to the statutory elements rather than merely reciting the six-factor list.

Below the criminal threshold, civil, administrative or professional measures may provide effective protection, but they are not inherently harmless. Administrative fines, licence restrictions and excessive damages can produce a chilling effect comparable to criminal punishment. Such measures require precise legal bases, independent decision-making, judicial review and proportionality. Civil remedies should prioritise cessation, correction, reply, declaratory relief and compensation linked to demonstrated harm rather than punitive awards imposed solely because expression is offensive.

Where speech remains lawful, non-coercive responses should be preferred: counter and alternative speech, education, media and digital literacy, support for targeted persons, professional ethics and inclusive public communication. This approach is consistent with the United Nations Strategy and Plan of Action on Hate Speech, which treats prevention and coordinated response as compatible with the protection of freedom of expression (United Nations, 2019), and with Armenian studies cautioning against reliance on criminalisation alone (Odabashyan, 2020; Ghazaryan and Asatryan, 2021). Public officials have an additional responsibility to avoid legitimising prejudice and to respond promptly to serious group-directed hostility without demanding unlawful censorship.

8.3. Stage Three: Governance of Dissemination

The third stage concerns dissemination governance. Public authorities and platforms should identify how recommendation, repetition, paid promotion,

coordinated behaviour and moderation decisions affected the scale and persistence of the risk. The central policy recommendation is not a universal notice-and-take-down duty, but a differentiated notice-and-action framework directed to actors with the technical and operational capacity to act on specific content. Internet access providers, domain registrars, hosting services and social-media platforms should not be treated as functionally interchangeable.

For platforms, a rights-based governance framework has six minimum components: clear differentiation between illegality and private rules; accessible and substantiated notice procedures; reasoned and contestable decisions; a proportionate ladder of interventions; systemic-risk assessment of recommender and advertising systems; and independent scrutiny through transparency reporting, audits and data access. These obligations should include attention to smaller languages and regional contexts, where moderation resources and appeal quality are often weaker.

For automated systems, the framework requires meaningful human control. Performance should be evaluated by language and affected group rather than only through aggregate accuracy. Political, journalistic, academic, artistic and human-rights documentation should receive enhanced contextual review. Automation may assist with triage, pattern detection and evidence preservation, but it cannot determine by itself whether a specific expression is protected, unlawful or criminally inciting.

8.4. Implementation, Evidence and Accountability

Implementation should be evidence-based. Public authorities should collect anonymised data on the legal provision used, protected characteristic, medium, procedural outcome, remedy and sanction. Platforms should report the ground for action, detection method, use of automation, type of restriction, appeal rate and reversal rate. Without disaggregated and comparable information, it is difficult to determine whether a framework protects targeted groups consistently, discriminates across languages, or is deployed selectively against political opposition and minority voices.

The Armenian case study illustrates the judicial dimension of the same framework. International threshold tests gain value when they discipline reasons and evidence, not when they function as labels. Their domestic application requires an explicit connection between each factor, the statutory offence and the sanction. A national judgment becomes internationally relevant not because it creates a global precedent, but because it reveals the practical conditions under which soft-law criteria can strengthen - or, if applied mechanically, weaken - rights-based adjudication.

9. CONCLUSION

The regulation of hate speech in digital environments must address two simultaneous risks: failure to protect persons and groups from discrimination, hostility and violence, and excessive restriction of democratic debate. Neither risk can be managed through lists of prohibited words, the speaker's self-description of intent, or the audience's subjective offence alone. The legally decisive inquiry combines context, speaker influence, intent, content and form, actual dissemination, and the likelihood of harm.

The Rabat six-factor test provides a valuable high-threshold structure. Its function is protective in both directions: it identifies exceptional cases in which criminal prohibition may be justified, and it constrains criminalisation where intent, extent or likelihood are not established. Recommendation CM/Rec(2022)16 complements that threshold with a graded framework of criminal, civil, administrative, professional, platform-based and non-legal responses. A broad policy definition of hate speech may coordinate these measures, but legal liability must remain attached to precise and foreseeable rules.

European digital regulation adds a procedural and systemic layer. The DSA does not harmonise the substantive law of hate speech or impose universal liability for user content. Its contribution lies in differentiated due-diligence duties: notice-and-action, reason-giving, redress, transparency, recommender-system governance, systemic-risk assessment, audit and data access. These mechanisms make platform power more accountable while preserving the distinction between illegal content, private community standards and lawful but harmful speech.

Algorithmic amplification may alter the extent and practical risk of expression, but it does not merge the responsibility of the speaker with the duties of the platform. Automated moderation should support, not replace, human legal judgment. The AI Act and the Council of Europe AI Framework Convention reinforce the importance of risk-sensitive transparency, oversight, accountability and equality, while the DSA and human-rights law remain the more direct frameworks for content-governance decisions.

The Armenian Court of Cassation's decision in Case No. ED/1485/01/21 is significant as an example of domestic judicial use of the Rabat factors in assessing public calls for violence. Its comparative value lies in structured contextual reasoning and in the tension, it exposes between a formal criminal offence and the continuing relevance of likelihood and proportionality. The judgment is not a general authority on platform liability or every form of group-based hate speech. Precisely that disciplined characterisation allows the case study to contribute to a broader international inquiry into how domestic courts can apply international standards while preserving legal certainty and freedom of expression.

The resulting governance model is therefore integrated but not undifferentiated. States must define liability precisely and provide effective remedies; courts must reason through the relevant threshold factors; platforms must govern dissemination transparently and proportionately; and automated systems must remain subject to meaningful human control. This division of responsibilities offers a more sustainable response than either unrestricted private censorship or exclusive reliance on criminal law.

References

1) Constitution of the Republic of Armenia (2015 amendments), art. 42.
2) Council of Europe, Committee of Ministers. (2022). Recommendation CM/Rec(2022)16 to member States on combating hate speech and explanatory memorandum.
3) Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms, ETS No. 5.
4) Council of Europe. (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225.
5) European Commission against Racism and Intolerance. (2015). General Policy Recommendation No. 15 on Combating Hate Speech.
6) European Union. (2008). Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. Official Journal of the European Union, L 328, 55–58.
7) European Union. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, L 277, 1–102.
8) European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.
9) Office of the United Nations High Commissioner for Human Rights. (2012). Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.
10) Republic of Armenia. (2021). Criminal Code, HO-199-N, arts. 329-330, as amended.
11) United Nations Human Rights Committee. (2011). General Comment No. 34: Article 19–Freedoms of opinion and expression, CCPR/C/GC/34.
12) United Nations. (1966). International Covenant on Civil and Political Rights, 999 UNTS 171.
13) United Nations. (2019). United Nations Strategy and Plan of Action on Hate Speech.
14) Behar and Gutman v. Bulgaria, no. 29335/13, European Court of Human Rights, 16 February 2021.
15) Beizaras and Levickas v. Lithuania, no. 41288/15, European Court of Human Rights, 14 January 2020.
16) Budinova and Chaprazov v. Bulgaria, no. 12567/13, European Court of Human Rights, 16 February 2021.

17) Court of Cassation of the Republic of Armenia, Criminal Chamber. (2026). Decision in Case No. ED/1485/01/21, 23 April 2026 (in Armenian).
18) Delfi AS v. Estonia [GC], no. 64569/09, European Court of Human Rights, 16 June 2015.
19) Handyside v. the United Kingdom, no. 5493/72, European Court of Human Rights, 7 December 1976.
20) Jersild v. Denmark, no. 15890/89, European Court of Human Rights, 23 September 1994.
21) Perinçek v. Switzerland [GC], no. 27510/08, European Court of Human Rights, 15 October 2015.
22) Sanchez v. France [GC], no. 45581/15, European Court of Human Rights, 15 May 2023.
23) Balavadze, M. (2025). Artificial intelligence and hate speech. Journal of Law, 2. https://doi.org/10.60131/jlaw.2.2025.10705
24) Brown, A. (2015). Hate Speech Law: A Philosophical Examination. Routledge. https://doi.org/10.4324/9781315714899
25) Ghazaryan, A., and Asatryan, D. (2021). Regulation of Hate Speech in the Legislation of the Republic of Armenia. Ara Ghazaryan Law Office (in Armenian).
26) Heinze, E. (2013). Hate speech and the normative foundations of regulation. International Journal of Law in Context, 9(4), 590–617. https://doi.org/10.1017/S1744552313000311
27) Khurana, U., Vermeulen, I., Nalisnick, E., van Noorloos, M., and Fokkens, A. (2022). Hate speech criteria: A modular approach to task-specific hate speech definitions. Proceedings of the Sixth Workshop on Online Abuse and Harms, 176-191. https://doi.org/10.18653/v1/2022.woah-1.17
28) Korre, K., Pavlopoulos, J., Gajo, P., and Barrón-Cedeño, A. (2024). Hate Speech According to the Law: An Analysis for Effective Detection. arXiv preprint arXiv:2412.06144. https://doi.org/10.48550/arXiv.2412.06144
29) Lepoutre, M. (2020). Hate speech laws: Expressive power is not the answer. Legal Theory, 25(4), 272–296. https://doi.org/10.1017/S135232522000004X
30) Odabashyan, H. (2020). Freedom of Expression and 'Hate Speech': Analysis of Armenian Legislation and International Experience. Luys Foundation (in Armenian).
31) Waldron, J. (2012). The Harm in Hate Speech. Harvard University Press.
32) Zufall, F., Hamacher, M., Kloppenborg, K., and Zesch, T. (2022). A legal approach to hate speech: Operationalizing the EU's legal framework against the expression of hatred as an NLP task. Proceedings of the Natural Legal Language Processing Workshop, 53-64. https://doi.org/10.18653/v1/2022.nllp-1.5