JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 22, No. 1 (2026)

From Hate Speech to Toxicity: depoliticization, algorithmic governance, and the transformation of harm in Digital Public Discourse

Roberto Bortone[a,1], Stefano Pasta[b]

[a] *National Office Against Discrimination - Rome (Italy)*

[b] *Catholic University of the Sacred Heart, Dept. of Education – Milan (Italy)*

(accepted: 20/5/2026; published: 23/5/2026)

Abstract

Over the last twenty-five years, hate speech has become a key category in international public policies, while digital environments have increasingly promoted the rise of the broader and more operational category of toxicity. This article argues that the shift from hate speech to toxic content should not be understood as a merely terminological substitution, but as a semantic and governmental transformation in the way discursive harm is identified, measured, and managed. The paper first reconstructs the historical and normative genealogy of hate speech, then examines the psychological, computational, and platform-based genealogy of toxicity, and finally compares the two frameworks through their conceptual, operational, and political implications. Particular attention is paid to the agency of platforms, algorithmic governance, content moderation, and the tension between discriminatory harm and conversational harm. The article suggests that toxicity offers scalability and technical operability, but may also contribute to the depoliticization of online harm if detached from histories of discrimination, protected characteristics, and asymmetries of power.

KEYWORDS: Hate Speech; Toxicity; Platform Governance; Content Moderation; Algorithmic Governance; Digital Public Sphere.

DOI
https://doi.org/10.20368/1971-8829/1136368

CITE AS
Bortone, R., & Pasta, S. (2026). From Hate Speech to Toxicity: depoliticization, algorithmic governance, and the transformation of harm in Digital Public Discourse. *Journal of e-Learning and Knowledge Society*, *22*(1).
https://doi.org/10.20368/1971-8829/1136368

1. Introduction

Over the last twenty-five years, "hate speech" has become a key concept in European and international public policies, without ever consolidating into a universally shared definition. In the lexicon of international institutions, it generally refers to forms of expression that incite, promote, disseminate, or justify hatred, discrimination, or violence against individuals or groups identified through protected characteristics. The Council of Europe Recommendation of 1997 provided an early broad and politico-normative framework, centered on expressions that "spread, incite, promote or justify" hatred based on intolerance (principle 1). The United Nations Strategy and Plan of Action on Hate Speech, in 2019, proposed a more descriptive and operational working definition, referring to communications that "attack" or use pejorative or discriminatory language against people or groups "on the basis of who they are" (p. 2). In 2022, Recommendation CM/Rec(2022)16 of the Council of Europe further refined the framework by introducing the semantically significant phrase "real or attributed" characteristics, thereby avoiding an implicit definition of the protected subject through the labels imposed by haters.

At the same time, however, digital environments have transformed the very nature of the phenomenon. Online hatred no longer appears only as direct insult or explicit incitement, but as a broader communicative ecosystem in which hate, harassment, disinformation, and other forms of abuse coexist and reinforce one another (Pasta, 2018; Bortone, 2023, 2025). In this context, the category of toxicity has gained growing relevance. Developed mainly in computational studies, content moderation, and platform policies, it serves to describe and measure communicatively harmful content even

[1] corresponding author - email: r.bortone@governo.it

when such content does not fall within the legal perimeter of hate speech. Empirical studies show, on the one hand, that online interactions tend to deteriorate into increasingly toxic exchanges when prolonged over time (Avalle et al., 2024; Pasta, 2022) and, on the other hand, that platform architectures can amplify anger and extreme content, encouraging the circulation of hatred and other forms of toxic communication (Munn, 2020).

This article interprets the shift from "hate speech" to "toxic language" not simply as a terminological replacement, but as a change in semantic regime and in the logic of governance. The analysis proceeds in three steps. First, it reconstructs the cultural and legal context in which the concept of hate speech emerged and stabilized. Second, it traces the psychological, computational, and platform-centered genealogy of toxicity. Finally, it compares the implications of the two lexicons in order to highlight their conceptual, operational, and political divergences.

2. Hate speech between rights, language, and social conflict

"Hate speech" is a historically constructed concept that matured within Western debate as a legal, political, and cultural category. If discriminatory and violent discourses, as well as the outcomes of what has been called the "hostile mind" (Santerini, 2021), have always existed, it is above all in the second half of the twentieth century that three decisive passages can be identified in the consolidation of the category.

A first turning point can be located in the post-war context, marked by the need to come to terms with the consequences of racist and antisemitic propaganda in Nazi-fascist totalitarian regimes (Santerini, 2005). In this framework, international law began to recognize that certain forms of expression are not neutral, but can contribute to the production and legitimation of violence. Documents such as the United Nations International Convention on the Elimination of All Forms of Racial Discrimination (1965) established the principle that incitement to hatred and discrimination is a legally and socially relevant problem, one that must be addressed not only through the regulation of actions, but also through attention to the social effects of language and the damage it can produce.

A second significant development took place in the United States between the 1970s and the 1990s, where debate on hate speech emerged within a structural tension between two fundamental principles: on the one hand, the very broad protection of freedom of expression guaranteed by the First Amendment; on the other, the demand - advanced especially by civil rights movements and university settings - to recognize the harm produced by racist, sexist, and homophobic language and acts, and to protect vulnerable groups from symbolic and material discrimination. In this genealogy, some Supreme Court decisions remain unavoidable reference points: Beauharnais v. Illinois (1952), often recalled for the issue of group defamation; Brandenburg v. Ohio (1969), which established the well-known standard of imminent lawless action; and R.A.V. v. City of St. Paul (1992), which made explicit how controversial it was to restrict bias-motivated expression. It is in this context that the term hate speech consolidated as an analytical category, even while remaining highly contested and normatively unstable. For this reason, Alexander Brown (2017) has convincingly argued that it should be treated not as a transparent and unitary concept, but as a heterogeneous family of expressive phenomena that remain historically and normatively disputed.

Still in the United States, but from a perspective very different from that of classical constitutional jurisprudence, Critical Race Theory introduced another decisive element: shifting the center of analysis from freedom of speech to the social effects of language. In this respect, *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (1993), by Mari Matsuda, Charles R. Lawrence III, Richard Delgado, and Kimberle Crenshaw, remains a key text. Although the authors also use the already established term "hate speech," they insist that racist language is not simply offensive, but can constitute a real-world injury in the form of exclusion and subordination. For this reason, they privilege the term "assaultive speech," which is less ambiguous because it focuses not on intentions or content alone, but on effects (Pasta, 2025). In this perspective, racist and other injurious speech acts represent a real form of harm that limits the freedom and equality of those targeted; for this reason, the authors argue, a democratic society may justify carefully tailored restrictions on speech. The text is especially significant because it anticipates several core features of the later development of the category of hate speech: the idea that formal freedom of expression can conceal substantive inequality, the attention to the lived experience of those who suffer racism and discrimination, the understanding of language as a practice of power and reproduction of hierarchies, the interdisciplinary nature of the inquiry, and the importance of intersectionality.

Alongside the U.S. debate, the European context developed a different approach, less absolutist in its understanding of freedom of expression and more oriented toward the protection of human dignity, the prevention of hateful discourse, and democratic coexistence, also in light of twentieth-century historical memory. In coherence with the differences between the European and American legal traditions (Ziccardi, 2016), European reflection adopted a more restrictive conception of freedom of expression when that freedom harms the dignity of others, treating language not only as individual manifestation but also as a social practice capable of producing exclusion, marginalization, and, in some cases, violence. In this framework, hate speech

consolidated as discourse that targets vulnerable groups. This is the context in which the Council of Europe adopted Recommendation No. R(97)20 in 1997, defining hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin." As an outcome of the previous trajectory, hatred is thus understood as a broad spectrum of expressions, as a social and structured phenomenon rather than a merely individual one, and as including less explicit forms such as aggressive nationalism, ethnocentrism, and hostility toward migrants - all correctly interpreted as fertile ground for more overt manifestations of hatred. It is not a binding legal norm, but rather a politico-normative definition intended to orient the policies of member states while leaving wide margins of interpretation and adaptation.

Recommendation CM/Rec(2022)16 of the Council of Europe later introduced several elements of novelty. First, it explicitly states that the characteristics on which hatred is based may be not only "real" but also "attributed," thereby recognizing that the target of hateful discourse is not objectively given but discursively constructed by the speaker. Second, the 2022 definition expands the range of relevant expressions by including not only incitement, promotion, or justification of hatred, but also forms of denigration and stigmatization, thus recognizing the continuity between explicit manifestations and subtler, everyday forms of discursive exclusion. Finally, the Recommendation strengthens attention to context and proportionality, underscoring the need for a graduated approach combining legal and non-legal measures.

In this perspective, hate speech progressively takes shape as a multifaceted concept: legal, because it is decisive in the balancing of freedom of expression and non-discrimination; political, because it calls into question the boundaries of the democratic environment and the conditions of participation; cultural, because certain utterances gradually become socially recognizable as degrading, threatening, or discriminatory. The history of the concept shows that we are not dealing with a merely descriptive category, but with an instrument built to define the nexus between language and inequality. From this point onward, the next transition becomes intelligible. Especially from a legal point of view, as public discourse increasingly moves into digital environments (Ziccardi, 2015; Bortone, 2023), the need emerges for a new law of the Internet (Rodota, 2004): the problem is no longer only how to prohibit or sanction certain expressions, but also the extent to which digital communication architectures contribute to defining, expanding, and normalizing them. It is not by chance that, as early as 1999, Lawrence Lessig argued in Code that in cyberspace the design choices embedded in software – the "code" – regulate behavior at least as much as legal rules do. This intuition now appears increasingly decisive for understanding the transformation of hate speech into a technical and infrastructural problem.

3. Beyond hate speech: the rise of toxicity

The lexicon of toxicity developed in a disciplinary context different from that of hate speech. While the latter emerged and consolidated within the language of human rights and non-discrimination, the notion of toxicity took shape mainly with the rise of the social web in the early 2000s, first within the psychology of communication and later in studies of computer-mediated communication and new digital environments.

A foundational reference is the work of the American psychologist John Suler, who in 2004 introduced the concept of the online disinhibition effect to describe the tendency of individuals to express themselves more freely in digital environments than in face-to-face interaction, often through more intense, impulsive, and aggressive forms of speech. Suler distinguishes between "benign disinhibition," associated with openness, self-disclosure, and emotional support, and "toxic disinhibition", marked by aggression, hostility, offensive language, and antisocial behavior (Suler, 2004).

It is important, however, not to move too quickly: toxic disinhibition does not yet coincide with toxic content. In the first case, we are dealing with a psychological and situational explanatory model that seeks to clarify why certain digital environments foster flaming, trolling, harassment, and other aggressive forms of conduct. In the second case, by contrast, we are already within a descriptive and classificatory logic, typical of platform moderation and computational research. Here, toxicity no longer indicates only a condition of production of discourse, but a quality of discourse itself: the degree to which content is perceived as rude, disrespectful, aggressive, or likely to deteriorate conversation. It is in this transition that toxicity becomes especially useful for annotation systems, benchmarks, and automated classification.

Unlike hate speech, toxicity does not necessarily presuppose a reference to vulnerable groups or protected characteristics. It describes instead a communicative quality: the degree to which content is perceived as hostile, corrosive, degrading, or potentially harmful to conversation. For this reason, the category of toxicity proved particularly suitable to the needs of platforms and computational research, both of which need to transform complex linguistic phenomena into annotatable, scoreable, and classifiable features. An emblematic case is Perspective API, launched in 2017 and developed by Jigsaw, a Google/Alphabet unit,

as a free machine-learning service designed to help platforms, publishers, and moderators anticipate how a comment might be perceived within a conversation. In its documentation, its main attribute – toxicity – is defined as a comment that is "rude, disrespectful, or unreasonable" and likely to discourage participation in the discussion. Around this infrastructure, an entire ecosystem of datasets annotated for toxicity, subtypes of toxicity, and identity references has been consolidated for the training and evaluation of automated models.

What emerges here is a passage from a socio-political notion of hatred to a technical-operational notion of "healthy conversation," one that can be handled through machine learning and inserted into a broader process of algorithmic governance of public discourse. The possibility of translating complex linguistic phenomena into scores, categories, and operational thresholds allows platforms to intervene at scale, but at the same time orients what is recognized as a problem. In other words, it is not only technology that "measures" toxicity; it is the need for measurement that selects which dimensions of discourse are made visible and treatable. In this sense, the category of toxicity is not neutral, but embedded in specific logics of governance that privilege what is computable over what is socially and historically situated.

The emergence of this concept is closely linked to the diffusion of the first online communication platforms - forums, chats, blogs - and to the need to understand phenomena such as flaming, trolling, and digital harassment (Pasta, 2018). Toxic disinhibition is thus interpreted as the effect of specific communicative conditions typical of the online environment, including anonymity, invisibility, emotional illiteracy, and a reduced perception of the social consequences of one's actions. The focus is not the content of discourse or its broader social dimension, but the conditions that facilitate the expression of behaviors that would otherwise remain inhibited.

Connected to the idea of toxic disinhibition is the concept of polarization, which has become central for describing the communicative dynamics of the social web and is often linked, in the age of platformization, to segmented public spheres and to the circulation of hateful discourse. In this direction, the work of Walter Quattrociocchi and his collaborators has progressively shifted attention from isolated content to the mechanisms of aggregation, mutual confirmation, and polarization that characterize digital information ecosystems: from echo chambers and confirmation bias to persistent patterns of toxicity across different platforms (Quattrociocchi & Vicini, 2013). Hate speech and toxic content, in this sense, can partially overlap, but both are understood as products of polarized information ecosystems – effects, rather than simple causes, of circulation dynamics such as echo chambers, homophily, infodemics, and polarization itself, including in their relation to algorithmic logics.

The shift from hate speech to toxicity does not necessarily entail an attenuation of the problem of hatred. On the one hand, it makes it possible to capture the gray area of discursive harms that are not always reducible to the most explicit forms of discriminatory hatred (Pasta, 2022). On the other hand, however, it moves the axis from law and responsibility toward metrics, policies, and risk management, thereby implicitly redefining the priorities of public and private intervention. The category of toxicity spreads also because it is more scalable, measurable, and comparable across multilingual and multicultural digital environments. Yet precisely this greater technical translatability can produce an effect of depoliticization: the centrality of asymmetries of power may be neutralized, and the focus can shift from justice to the management of "harmful" or "unsafe" communication across a broad range of behaviors.

4. Comparing hate speech, toxicity, and platform definitions

At this point, it becomes useful to compare three definitional families that frame current interventions against processes of digital targeting: the classical definitions of hate speech elaborated by international institutions; the definitions of toxicity developed in the psychology of online communication and later in computational research; and the operational definitions adopted by digital platforms for moderation and enforcement, as of May 2026.

The tables below do not aim to provide an exhaustive repertory, but rather a reasoned comparative grid gathering the most influential definitions. For the hate speech column, a selection of particularly influential institutional formulations within the European and international context has been included; for the toxicity column, definitions or formulations that have played a significant role within the psychological tradition and in the subsequent computational translation of the concept; and for the platform-related column, official policies that clearly illustrate the transformation of a socio-political category into a moderation taxonomy.

The criteria adopted for the comparison are: disciplinary matrix of origin, primary object, necessity of reference to protected characteristics, type of presumed harm, threshold of problematicity, role of context, and regulatory output. As for the selected platforms, Meta, YouTube, TikTok, Twitch, and Discord do not exhaust the landscape, but represent five different ecologies of governance: generalist social networking, video sharing, short-video platforms, live streaming, and community chat/server environments.

Table 1 summarizes the principal differences among three definitional families that overlap only partially.

Table 1 - Criteria for comparing hate speech, toxicity, and digital platform definitions.

Criterion	Hate speech	Toxicity	Operational platform definitions
Originating framework	Human rights, anti-discrimination, international case law	Psychology of online communication, natural language processing, content moderation studies	Trust & safety, community standards
Primary object	Expressions that incite, promote, disseminate, or justify hatred, violence, or discrimination; or that denigrate persons or groups	Comments or content perceived as rude, aggressive, disrespectful, or likely to deteriorate conversation	Attacks, threats, dehumanization, degradation, slurs, incitement to hatred or violence according to proprietary taxonomies
Centrality of reference to protected characteristics	Yes, central	No, not necessarily	Generally central in hate policies; often integrated with broader policies on harassment and abusive behavior
Typical target	Groups or members of vulnerable or inferiorized groups	Generic interlocutors or participants in discussion	Individuals or groups, with differentiated protections for protected attributes
Type of harm presupposed	Discriminatory harm, violation of dignity, exclusion, threat to democratic coexistence	Deterioration of the discursive climate, interactional hostility, abandonment of discussion	Risk to user safety, service integrity, regulatory compliance, reputational harm
Threshold of problematicity	Hatred, discrimination, violence, denigration; often graduated	Rudeness, disrespect, aggression, likelihood of derailing discussion	Operational thresholds such as “direct attack”, “promotes violence or hatred”, “degrades”, or “dehumanizes”
Role of context	High: history, relations of power, social meaning, asymmetries	Weaker or more standardized, especially in automated models	Relevant, but filtered through internal guidelines, exceptions, and scalability requirements
Technical operability	Medium-low	High	Very high
Typical output	Norms, recommendations, case law, criminal/civil/ administrative sanctions	Scores, benchmarks, datasets, auditing, risk assessment	Removal, downranking, strikes, demonetization, labeling, limitation of reach
Main risk	Ambiguity or definitional overbreadth	Depoliticization of discriminatory harm; model bias	Reduction of a socio-political category to a problem manageable at scale

As the table shows, classical hate speech serves to name discriminatory harm; toxicity serves to make communicative harm operationally treatable; and platforms produce a hybrid category that preserves the reference to protected characteristics while bending it toward the needs of enforcement and private governance. In this way, platform policies transform a socio-political category into a taxonomy of moderation

The comparison suggests that classical hate speech definitions are semantically denser. They do not merely say that content is offensive or aggressive: they place it within a framework of discrimination, dignity, potential violence, and democratic coexistence. This density makes them theoretically richer, but also less easily translatable into automated criteria. By contrast, toxicity is thinner on the political plane but much stronger operationally: it is easier to annotate, quantify, and model.

In this framework, platform policies reveal a precise conceptual transformation: the socio-political category of hate is retained only insofar as it can be divided into moderable acts, operational thresholds, exceptions, and enforcement levels. The real fracture, therefore, is not simply between “hate” and “toxicity,” but between discriminatory harm and conversational harm. Hate speech primarily protects against subordination and exclusion; toxicity primarily protects against deterioration of conversation and hostile climates; platforms, finally, also protect themselves, since service integrity, perceived safety, governability of space, and regulatory compliance are structural parts of their definitions.

5. Platform agency: the technical reformulation of hate speech

In conditions of persistent fragmentation of the transnational normative framework governing the web, digital platforms and their operating architectures now constitute an infrastructural level that not only organizes the circulation of content, but actively intervenes in its definition, classification, and possible removal. If hate speech allows us to analyze the content

Table 2 - Criteria for comparing hate speech, toxicity, and digital platform definitions.

Hate speech	Toxicity	Platforms
Council of Europe 1997: "all forms of expression which spread, incite, promote or justify hatred based on intolerance"	Suler 2004: online disinhibition helps explain why people express themselves online in more intense, impulsive, or aggressive ways	Meta: "direct attacks against people ... on the basis of protected characteristics"
Council of Europe 2022: expressions that incite, promote, disseminate, or justify violence, hatred, or discrimination, or denigrate, on the basis of real or attributed characteristics	Perspective API (since 2017) / computational tradition: toxicity as an annotatable property of comments, useful for estimating how rude, disrespectful, or conversation-damaging they may be	YouTube: content that "promotes violence or hatred" against individuals or groups on the basis of protected status
United Nations 2019: communication that attacks or uses pejorative or discriminatory language toward persons or groups "on the basis of who they are"	Perspective subtypes: severe toxicity, identity attack, insult, and other subcategories that decompose communicative harm into classifiable attributes	TikTok: explicit or implicit content attacking a protected group; dehumanization, supremacy, segregation, and discrimination are also prohibited
EU Framework Decision 2008/913/JHA: public incitement to violence or hatred against a group or a member of a group defined by race, colour, religion, descent, or national or ethnic origin	NLP / moderation research: a metric and generalizable category, useful for benchmarks, auditing, and risk assessment beyond the legal core of hate speech	Twitch: content or behavior that discriminates, denigrates, harasses, or encourages violence based on protected characteristics
Contemporary European approach: a graduated concept, to be addressed through criminal, civil, administrative, and non-legal measures proportionate to gravity	Contemporary line of inquiry: toxicity is often treated as an indicator of deterioration of discursive space and of the probability of making conversation unlivable	Discord: expressions that degrade, vilify, or dehumanize, incite hostility, or promote harm on the basis of protected characteristics

and socio-political dimension of discourse, and toxic disinhibition explains some of its psychological conditions of production, platforms – as non-neutral spaces of mediation – are real actors exercising definitional power over public discourse. By operationally reformulating the problem of hate speech through their own policies, they exercise not only regulatory power, but also genuinely epistemic power: they select the categories through which the phenomenon becomes knowable, measurable, and governable.

This conceptual transformation can be read as a form of privatization of the regulation of public discourse, in which global economic actors assume functions historically closer to law and democratic institutions. By privileging an approach centered on the removal or downgrading of explicitly aggressive content, platforms risk neglecting subtler, more diffuse, and normalized forms of hatred, operating through narratives, stereotypes, and discursive constructions that are less visible but no less socially effective. Recent regulatory and judicial developments also show that online harm is increasingly being read as an effect of the environment designed by the platform, and not only of single pieces of content.

Shifting attention from structures of meaning to the more superficial manifestations of discourse, as platforms tend to do, brings out a tension between the needs of scalability and the need for critical understanding. In a previous analysis based on a corpus of cases of online racism, it was argued that there is a "return of race in the absence of scientific credibility," a phenomenon that can be reread in light of the transition from hate speech to toxic content: "A concept of 'race' defeated by science is thus affirmed, yet socially accepted by our implicit popular pedagogies and therefore able to re-emerge in the collective consciousness even without scientific credibility" (Pasta, 2019, pp. 185-186). In a similar way, within the communicative flow of the social web, racist discourse – loaded with history and meaning – may be emptied of meaning and thereby rendered socially acceptable.

This transition also finds empirical confirmation. Within Hate Studies, research aimed at detecting forms of hatred can be divided between studies that rely exclusively on machine-learning methods and those that combine automated retrieval with human classification (Forzinetti et al., 2024; Pasta, 2024a). As shown by studies conducted by the Mediavox Observatory at Università Cattolica on antisemitism (Pasta, 2024b; Pasta et al., 2021), Islamophobia (Pasta, 2023a), and anti-Gypsyism (Pasta, 2023b), systems based solely on automated detection - for example through dictionaries of offensive words or linguistic patterns - return results that differ significantly from context-sensitive human annotation when the same corpus is subjected to the two methods and evaluated through a confusion matrix.

What is at stake, therefore, is the protagonism of powerful private actors in the digital sphere against the

relative weakness of supranational actors that – especially in Europe – have attempted to regulate platform responsibility, sometimes ending in open conflict (Human Rights Watch, 2023). It is important to recall here the connection between the genesis of hate speech and the perspective of human rights, at a historical moment in which supranational law itself is being challenged by war, by the crisis of liberal democracy, and by claims voiced from within the global web elite. Donatella Di Cesare (2025) has spoken of "technofascism" to describe the convergence of technocratic and ethnocratic tendencies within parts of the new right. In a different but convergent register centered on the primacy of technique and security, Alexander C. Karp (CEO of Palantir Technologies) and Nicholas W. Zamiska, in The Technological Republic (2025), call for a renewed alliance between technological industry, state power, and national security. More than reconstructing the genesis of the concept of hate speech, these texts help illuminate the ideological context in which the boundaries of acceptable discourse online are being redrawn today.

6. Between hate speech and toxic content: some consequences

The comparison developed in this article shows that the passage from hate speech to toxicity concerns the way in which discursive harm is named, measured, and governed. Hate speech retains a normative force that is difficult to replace, because it allows language to be read within relations of power, histories of subordination, and regimes of inferiorization that exceed mere conversational aggression, especially in relation to vulnerable groups (Pasta, 2018). Toxicity, by contrast, has established itself because it offers a flexible and technically manageable operational category, functional to the measurement, classification, and management of communicative risk in digital environments, and capable of capturing a gray area of discursive harm that does not always coincide with the hard core of discriminatory hate. This gray area of communicative harm is not entirely alien even to European anti-discrimination law: both Directive 2000/43/EC and Directive 2000/78/EC qualify as discrimination unwanted conduct that violates a person's dignity and creates an intimidating, hostile, degrading, humiliating, or offensive environment. This shows that the problem of environmental and relational harm partly predates the computational category of toxicity, even if it does not coincide with it.

The point, then, is not to establish in the abstract which of the two categories should prevail, but to recognize that both refer today to a broader transformation of the digitalized public sphere (Bortone, 2023), in which the governance of discourse is increasingly entrusted to technical infrastructures, risk metrics, and criteria of optimization (Habermas, 2005; van Dijck et al., 2018). The Digital Services Act, for its part, requires very large online platforms (VLOPs) to identify and mitigate systemic risks deriving not only from illegal content, but also from service design, recommender systems, automated moderation, and targeted advertising. This represents an important attempt to respond to the platform agency discussed above.

A first methodological consequence concerns detection. Those who continue to work with the category of hate speech – whether in detection, in evaluating the impact of offensive content on victims, in building policy, or in educational research – can no longer ignore the spread of the category of toxicity, which now informs a substantial part of moderation practices, risk assessment, and computational research. Conversely, those who employ toxicity should make explicit why they do so, and to what extent that choice entails a shift from discriminatory harm to conversational harm. Content that is socially significant from the standpoint of discrimination may appear only weakly "toxic" according to computational parameters, while highly conflictual but non-discriminatory content may be classified as problematic. The result is a possible divergence between what is technically detectable and what is socially meaningful.

A second consequence concerns the fact that platforms do not merely moderate hate speech: they contribute to reformulating it. The risk, as this contribution has sought to show, is that the political dimension of hateful discourse – understood as the expression of hierarchies, conflicts, and inequalities – is progressively recoded as an issue of compliance with communicative standards, service safety, or quality of user experience. Depoliticization would therefore not consist in the disappearance of the problem, but in its reformulation in terms compatible with the governability of the platform. In moderation practices, platforms tend to privilege hybrid categories that combine elements of classical hate speech with logics proper to toxicity. This undoubtedly allows greater operational effectiveness - in removal, downranking, or visibility limitation - but also reformulates the problem itself: hate speech is increasingly treated as an issue of individual communicative behavior, linked to the protection of digital well-being, rather than as an expression of structures of social inequality and injustice.

A third consequence concerns public policy. The shift from hate speech to toxicity raises significant questions in terms of accountability and governance. If hate speech has traditionally been anchored to legal frameworks and to systems of rights protection, toxicity occupies a more fluid space, often governed by private policies and technical standards. In the absence of a fully shared transnational normative framework, this implies a displacement of power toward platforms, which not only apply rules but also help define them by

selecting which dimensions of the phenomenon are made visible and governable.

A fourth consequence concerns education. If one limits oneself to addressing "toxicity" as a problem of communicative behavior - for example by promoting respectful communication or conflict-management skills - one risks overlooking the structural dimension of hate speech, linked to stereotypes, prejudice, power relations, and socially situated practices of inferiorization. By contrast, a critically informed educational approach should hold together both dimensions: it should develop communicative and relational skills for onlife citizenship (Pasta & Rivoltella, 2022) while also promoting a critical literacy capable of recognizing the nexus between language, inequality, and the social construction of difference (Pasta, 2023c; Faloppa, 2026). The challenge, then, is to construct integrated definitional models capable of combining the technical precision of automated detection with the interpretive depth of the social sciences and law, so that the complexity of the phenomenon is not reduced to what is merely measurable.

A further element reinforces this need for conceptual caution. Recent regulatory and judicial developments show that digital harm is increasingly being read as the effect of the environment designed by the platform, and not only of the individual content item. In the United States, Section 230 of the Communications Decency Act (1996) has historically shielded intermediaries from classical editorial liability. The Supreme Court did not substantially curtail this framework in Gonzalez v. Google (2023), and Twitter, Inc. v. Taamneh (2023); and in Moody v. NetChoice (2024), it recognized broad editorial discretion in content moderation. At the same time, the March 2026 verdict in K.G.M. v. Meta Platforms, Inc., et al. shifted attention toward the design and functioning of Instagram and YouTube – that is, toward the design of the environment rather than only third-party content – while, at the European level, the Commission preliminarily found Meta in breach of the DSA for failing effectively to prevent access by users under the age of 13. All this does not coincide with the category of toxicity, but it makes more plausible the idea that digital harm must also be read as an effect of socio-technical architectures, recommender systems, and logics of engagement.

Finally, a critical piece of evidence deserves to be signaled as a future development of research: the possibility that a new platform environment is emerging. Some evidence suggests that platforms do not merely remove clearly hateful or violative content, but may also reduce the visibility of politically sensitive content, content difficult to monetize, controversial content, or content deemed risky for the safety and governability of the environment. A Human Rights Watch report of 2023 documented more than 1,050 cases of removal or undue suppression of Palestinian or pro-Palestinian content on Facebook and Instagram; Meta announced in 2024 that it would no longer proactively recommend political content on Instagram and Threads, and from 2025 it discontinued political, electoral, and social issue ads in the European Union. In this context, even content oriented toward peace or civic testimony may enter a zone of reduced sponsorizability or reduced visibility, not because it is assimilated to hate, but because it is classified as politically sensitive or as social-issue content. A report gathered by the authors from an association that attempted to promote a peace event points in the same direction, although it has no generalizable probative value. The provocative question – can even peace be "toxic"? – should therefore be understood not literally, but as a critical question: in an environment optimized for brand safety, risk management, and governability, even pacifist content, war testimonies, citizen journalism, or digital activism may become problematic for the platform, although they are neither hateful nor illegal.

For this reason, one possible line of future research would be to verify whether the feeds of the main social media platforms have been progressively "cleaned" not only of explicitly hateful content, but also of images, testimonies, and forms of digital activism that do not propagate hatred and yet make the environment less marketable, more conflictual, or more difficult to govern according to platform parameters.

Acknowledgements

The present contribution is the joint work of the two authors; however, it should be noted that §§ 1, 4, 6 were written by Roberto Bortone and §§ 2, 3, 5 by Stefano Pasta.

References

Avalle, M., Di Marco, N., Etta, G. et al. (2024). Persistent interaction patterns across social media platforms and over time, Nature, 628, 582-589, https://doi.org/10.1038/s41586-024-07229-y.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification, Companion Proceedings of the 2019 World Wide Web Conference ACM, https://doi.org/10.1145/3308560.3317593.

Bortone, R. (2023). Molto social, troppo dark. Tra hate speech, propaganda, metaverso e intelligenza artificiale: i rischi del web oggi, Roma, Fefè.

Bortone, R., & Pistecchia, A. (2025, Eds.). L'ultimo pregiudizio. L'antiziganismo tra storia e attualità, Roma, Nuova Cultura.

Brown, A. (2017). What is Hate Speech? Part 1: The Myth of Hate. Law and Philosophy, 36(4), 419-468, https://doi.org/10.1007/s10982-017-9297-1.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language, Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515, https://doi.org/10.1609/icwsm.v11i1.14955.

Di Cesare, D. (2025). Tecnofascismo. Torino, Einaudi.

Faloppa F. (2020), #Odio. Manuale di resistenza alla violenza delle parole, Torino, UTET.

Faloppa, F. (2026). Disarmare il discorso. Sulla militarizzazione del linguaggio. Firenze, Effequ.

Forzinetti, E., Della Vedova, M., Pasta, S., & Santerini, M. (2024). Indicators for characterising online hate speech and its automatic detection (arXiv:2402.08462), Cornell University, http://arxiv.org/abs/2402.08462.

Gelber, K. (2021). Differentiating hate speech: a systemic discrimination approach, Critical Review of International Social and Political Philosophy, 24(4), 393-414, https://doi.org/10.1080/13698230.2019.1576006.

Habermas, J. (2005). Storia e critica dell'opinione pubblica. Roma-Bari, Laterza.

Karp, A.C., & Zamiska, N.W. (2025). The Technological Republic: Hard Power, Soft Belief, and the Future of the West, New York, Crown Currency.

Lessig, L. (1999). Code and Other Laws of Cyberspace. New York, Basic Books.

Matsuda, M., Lawrence III, C.R., Delgado, R., & Crenshaw, K. (1993). Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment. New York, Routledge.

Munn, L. (2020). Angry by design: toxic communication and technical architectures. Humanities and Social Sciences Communications, 7, 53. https://doi.org/10.1057/s41599-020-00550-7.

Pasta, S. (2018). Razzismi 2.0. Analisi socio-educativa dell'odio online. Brescia, Morcelliana Scholé.

Pasta, S. (2019). Razzismi espliciti banalizzati. L'ambiente digitale e il 'ritorno della razza'", in Santerini M. (ed.), *Il nemico innocente. L'incitamento all'odio nell'Europa contemporanea. The Innocent Enemy. Hate incitement in contemporary Europe*, Milano, Guerini e Associati, 173-190.

Pasta, S. (2022). Social network conversations with young authors of online hate speech against migrants, in Monnier A., Boursier A., Seoane A (Eds.), Cyberhate in the Context of Migrations, London, Palgrave MacMillan, 187-214.

Pasta, S. (2023a). Discours de haine en ligne. Une analyse des tweets islamophobes entre automatismes et évaluation qualitative, in Karkun A., Jovelin E. (Eds.), *Vers une paix durable. Une perspective interculturelle*, Paris, Éditions du Cygne, 217-237.

Pasta, S. (2023b). Hate Speech Research: Algorithmic and Qualitative Evaluations. A Case Study of Anti-Gypsy Hate on Twitter, *REM. Research on Education and Media*, 15(1), 130-139.

Pasta, S. (2023). Tackling online hate speech with the involvement of targeted groups. The methodological proposal of the project REASON – REAct in the Struggle against ONline hate speech, QTimes. Journal of Education, Technology and Social Studies, XV(3), 429-445, doi: 10.14668/QTimes_15330.

Pasta, S. (2024a). Lo "spettro dell'odio online": una proposta di classificazione tra valutazioni algoritmiche e qualitative, in Crescenza G. (eds.), *Educare in tempi di odio e violenza. Sfide pedagogiche e istituzionali*, Bari, Progedit, 113-126.

Pasta, S. (2024b). Hate Studies tra logica computazionale e classificazione umana. Un caso studio sull'antisemitismo in Twitter, Scholé. Rivista di educazione e studi culturali, LXII (1), 230-252.

Pasta, S. (2025), *Countering Hate Speech in the Postdigital: A Challenge for 'Onlife Citizenship'*, in Gomez Paloma F., di Tore P.A., Mangione G.R.J. (eds.), *Teacher Training and Student Learning – Past Values, Present Uncertainties and Future Prospects*, IntechOpen, London, pp. 155-174, DOI: 10.5772/intechopen.1011481.

Pasta, S., Santerini, M., Forzinetti, E., & Della Vedova, M. (2021). Antisemitism and Covid-19 on Twitter. The search for hatred online between automatisms and qualitative evaluation, Form@re. Open Journal per formazione in rete, XXI, 3, 288-304,

Pasta, S., Rivoltella, P.C. (2022, Eds.), Crescere onlife. L'Educazione civica digitale progettata da 74 insegnanti-autori, Brescia, Scholé.

Quattrociocchi, W., & Vicini, A. (2023). Polarizzazioni. Informazioni, opinioni e altri demoni nell'infosfera, Milano, FrancoAngeli.

Rodota, S. (2004). Tecnopolitica. La democrazia e le nuove tecnologie della comunicazione, Roma-Bari, Laterza.

Santerini, M. (2005). Antisemitismo senza memoria, Roma, Carocci.

Santerini, M. (2021). La mente ostile, Milano, Raffaello Cortina.

Sellars, A. (2016). Defining Hate Speech, Berkman Klein Center Research Publication No. 2016-20, https://cyber.harvard.edu/publications/2016/DefiningHateSpeech.

Suler, J. (2004). The Online Disinhibition Effect. CyberPsychology & Behavior, 7(3), 321-326, https://doi.org/10.1089/1094931041291295.

van Dijck, J., Poell, T., & de Waal, M. (2018). The Platform Society: Public Values in a Connective World, Oxford, Oxford University Press, https://doi.org/10.1093/oso/9780190889760.001.0001

Waldron, J. (2012). The Harm in Hate Speech. Cambridge, Harvard University Press.

Ziccardi, G. (2015). Internet, controllo e libertà Trasparenza, sorveglianza e segreto nell'era tecnologica. Milano, Raffello Cortina.

Ziccardi, G. (2016). L'odio online Violenza verbale e ossessioni in rete. Milano, Raffello Cortina.

Normative and institutional documents

Council of Europe, Committee of Ministers (1997), Recommendation No. R(97)20 on "Hate Speech", Strasbourg.

Council of Europe, Committee of Ministers (2022), Recommendation CM/Rec(2022)16 on Combating Hate Speech, Strasbourg.

Council of Europe. (2022), Explanatory Memorandum to Recommendation CM/Rec(2022)16 on combating hate speech, Strasbourg.

Council of the European Union (2000), Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

Council of the European Union (2000), Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.

Council of the European Union (2008), Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

European Parliament and Council (2022), Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act).

United Nations (1965), International Convention on the Elimination of All Forms of Racial Discrimination.

United Nations (2019), United Nations Strategy and Plan of Action on Hate Speech, New York.

Policy documents, case law, and essential web sources

Cornell Law School, Legal Information Institute, 47 U.S. Code § 230 - Protection for private blocking and screening of offensive material.

Discord (2025), Hateful Conduct Policy Explainer, https://discord.com/safety/hateful-conduct-policy-explainer.

European Commission (2026, 29 April), Commission preliminarily finds Meta in breach of Digital Services Act for failing to prevent minors under 13 from using Instagram and Facebook.

Gonzalez v. Google LLC, 598 U.S. (2023).

Human Rights Watch (2023), Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook, https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-on-instagram-and-facebook

Jigsaw. Supporting Online Conversations at Scale with AI, https://jigsaw.google/our-work/supporting-online-conversations/.

K.G.M. v. Meta Platforms, Inc., et al. (California jury verdict, March 2026).

Meta (2024), Actualizacion sobre nuestro abordaje al contenido politico en Instagram y Threads, https://about.fb.com/ltam/news/2024/02/actualizacion-sobre-nuestro-abordaje-al-contenido-politico-en-instagram-y-threads/.

Meta (2025), Ending Political, Electoral and Social Issue Advertising in the EU in Response to Incoming European Regulation. https://about.fb.com/news/2025/07/ending-political-electoral-and-social-issue-advertising-in-the-eu,

Moody v. NetChoice, LLC, 603 U.S. (2024).

TikTok (2025), Countering Hate Speech & Behavior. https://www.tiktok.com/safety/en/countering-hate/.

Twitch (2025), Hateful Conduct Education Resource, https://www.twitch.tv/p/safety/education-portal/en/articles/hateful-conduct/.

Twitter, Inc. v. Taamneh, 598 U.S. (2023).

YouTube (2025), Hate speech policy, https://support.google.com/youtube/answer/2801939.