VISUAL REVIEW | Vol. 18, No. 9s, 2026 | ISSN 2695-9631

International Visual Culture Review / Revista Internacional de Cultura Visual

HUMAN RESOURCES MANAGEMENT IN SOCIAL MEDIA CONTENT MODERATION

Perspectives from the Digital Services Act

Gestión de recursos humanos en la moderación de contenidos de redes sociales. Perspectivas desde la Ley de Servicios Digitales

MARIAN NÚÑEZ CANSADO
ORCID: 0000-0002-6658-3996 · mariaangeles.nunez@uva.es

ÁNGEL QUINTANA GÓMEZ
ORCID: 0000-0002-5139-8755 · angel.quintana@ulpgc.es

KEYWORDS

Content moderation
Social media
Digital Services Act (DSA)
Communication ethics
European Union (EU)
Communication
Advertising

ABSTRACT

This study evaluates the implementation of human resources in content moderation in large digital platforms (VLOPs) within the European Union, using a quantitative methodology based on documentary and data analysis. The results show significant disparities in the ratio of moderators per language, with TikTok and YouTube leading in coverage, while LinkedIn and X present critical gaps. It is concluded that, although the platforms comply with regulatory dictates, the lack of full language coverage and inadequate working conditions pose significant risks to vulnerable communities and moderators.

PALABRAS CLAVE

Moderación de contenidos
Redes Sociales
Ley de servicios digitales
Ética de la comunicación
Unión Europea
Comunicación
Publicidad

RESUMEN

Este estudio evalúa la implementación de recursos humanos en la moderación de contenidos en grandes plataformas digitales (VLOPs) dentro de la Unión Europea, utilizando una metodología cuantitativa basada en análisis documental y de datos. Los resultados muestran disparidades significativas en la ratio de moderadores por idioma, con TikTok y YouTube liderando en cobertura, mientras que LinkedIn y X presentan vacíos críticos. Se concluye que, aunque las plataformas cumplen con los dictámenes regulatorios, la falta de cobertura lingüística completa y las condiciones laborales inadecuadas representan riesgos significativos para comunidades vulnerables y moderadores.

1. Introduction

The ability to store and share online content on a large scale has significantly transformed the way people relate to one another, giving such content a crucial role in personal communication, identity construction and cultural expression (Manovich, 2020; Mitchell, 2017). However, this torrent of content raises major ethical and privacy dilemmas, particularly with regard to the management of digital images—in aspects such as the selection, organization and display of images—as well as their potentially violent and/or hate-inciting nature.

Editing content in digital media gives rise to dilemmas relating to copyright and the consent of the people photographed, as well as to the handling of personal data that may be involved in the content used, among other issues. It is important to note that, although social media allow users to share content quickly and on a massive scale, these platforms include clauses in their terms and conditions that grant the social network a broad license over the content published by users. This contractual mechanism, as Hernández García (2021) states, means that users assign rights of use to the platform, allowing it to reuse their content within the limits defined in the terms of use. A substantial body of scientific literature highlights the problems of unauthorized use of content and the violation of individual privacy, which can often compromise the safety of social media users (Nissenbaum, 2011). Alongside these problems, it should be stressed that reuse and lack of context when content is disseminated on digital platforms can lead to plagiarism and distorted representation, which may negatively affect public perceptions of certain individuals and/or social groups (Boyd & Marwick, 2017). Finally, we must not overlook content that promotes violence, incites hatred or involves child exploitation, which can have a profoundly harmful impact on the construction of both individual and collective identity.

It is important to recognize that this situation is especially sensitive in the case of children, who form a group that is particularly vulnerable to the implications of digital content dissemination (Livingstone & Third, 2017; Lievens et al., 2019). Some of these risks may become apparent, for example, in the digital exploitation of image and identity, which may lead to the development of a distorted self-concept that gives rise to serious psychosocial problems (Núñez-Cansado et al., 2021).

To the many risks associated with social media content and the right to privacy, it is essential to add the impact of the algorithms used by the platforms. These systems, designed to maximize user interaction and time spent on the platform, tend to prioritize sensationalist, polarizing or emotionally charged content. In 2021, a former employee of Facebook-Meta called the company's management into question by leaking dozens of documents to *The Wall Street Journal*. These documents revealed questionable practices that fostered a negative influence on the youngest population, stemming from the content and algorithms implemented by the platform. As the employee stated in her testimony before the European Parliament:

"The algorithmic models used by social networks such as Instagram were specifically designed to artificially encourage social comparison, eroding young women's self-perception of their bodies, social practices and financial resources" (Jiménez González & Cancela Rodríguez, 2023, p. 91).

Arturo Béjar, head of Facebook's '*Integrity and Care*' department, revealed in the documentary Redes sociales: la fábrica del terror [Social Media: The Terror Factory] that the platform was aware of, and ignored, the harm caused to its users. In June 2020, the former executive sent his now well-known email to Mark Zuckerberg and the rest of the company's senior management, providing chilling data—for instance, that 51% of Instagram users reported negative content and that 21.8% of teenagers aged 13 to 15 reported having experienced direct harassment—and stating categorically that Instagram was probably responsible for the largest-scale sexual harassment in human history (Évole & Lara, 2024).

This brought to light a reality that has sparked a new debate on the dangers of social media and created the need to put in place systems to regulate the content disseminated by platforms.

Although legal provisions on the use and protection of data on the internet already existed—in particular, Europe's General Data Protection Regulation (GDPR)—these regulations proved insufficient in the face of the harm reported. Germany therefore enacted the Network Enforcement Act, better known as NetzDG, in 2017, which served as the basis for various provisions such as the EU Code of Practice on Disinformation, adopted by the European Union in 2018, and the ministerial order approved by Spain's National Security Council on 30 October 2020, which established the Procedure for Action against Disinformation. UNESCO joined in this concern and held a Seminar on Content Moderation and Freedom of Expression on Social Media, which addressed content regulation and the protection of human rights, exploring issues such as procedures (AI vs. human resources), moderation criteria, transparency and the need for large platforms to be accountable to the relevant authorities.

The absence of specific regulation, and the challenge posed by the need to moderate digital content in order to mitigate the associated risks, compelled corporations and governments to make commitments, which materialized in the European Union's enactment of the Digital Services Act (DSA), which would enter into force on 25 August 2023 (European Union, 2022). The DSA is presented as a key instrument for establishing clear rules that guarantee a safe digital environment. The act lays down specific obligations that companies must respect, such as algorithmic transparency, a ban on intrusive and misleading advertising practices, a strengthened reporting mechanism and the removal of illegal content; it also introduces measures to protect users, especially minors, from harmful content. Under the regulation, very large online platforms (VLOPs) with more than 45 million active users in the European Union must take additional measures to manage the systemic risks arising from their activity and are required to be accountable through periodic reports on content moderation and political advertising. The fundamental aim of the act is to balance safety, transparency and accountability in the European Union's digital ecosystem, making platforms responsible for the content they host and protecting users from harmful content, harmful algorithms and the abusive control of personal data.

However, these laws still face major challenges in their implementation and enforcement in a context of constant digital transformation, where content moderation often proves complex, whether because of conflicts with commercial interests or the absence of specific regulations in certain virtual spaces (Jiménez González & Cancela Rodríguez, 2023).

2. Content moderators

The risks that the misuse of social media content can entail have increased the need not only for an in-depth analysis of legislation, but also for commitment on the part of social media companies. The variety of content, recommendation algorithms and social graphs have given rise to numerous complaints and criticisms of content platforms, which has forced media companies to commit to constant and rigorous monitoring on behalf of their users, posing both logistical and public relations challenges (Gillespie, 2020). This is how "commercial content moderation", or "platform moderation", has become so vitally important. In recent years there has been a sharp increase in scientific literature addressing the main issues surrounding this new professional profile, ranging from concerns about working conditions and the analysis of legislation to democratic legitimacy and even the lack of transparency and accountability of large corporations (Gorwa et al., 2020).

Content moderation is the organized practice of screening user-generated content posted on internet sites (Roberts, 2019), and has been defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann, 2017: 18). Its aim is to ensure a safe and respectful environment for all social media users by preventing the spread of inappropriate, illegal or harmful content, such as hate speech, violence, disinformation, harassment or explicit material. Material may be reviewed before it is distributed, or it may be assessed following a complaint from a user, a site administrator, an affected company, and so on.

Over the last year, this new professional profile has prompted debate on three fundamental issues: the creation of basic manuals for the practice of the profession; the recognition of

psychosocial risks and, consequently, the recognition of their effects as occupational diseases; and the rules governing the coexistence of human regulation and AI regulation.

In 2022, the European Agency for Safety and Health at Work (EU-OSHA) published the report Occupational safety and health risks of online content review work provided through digital labour platforms, which identified exposure to violent content, crime, abuse and illegal content as a source of stress that could cause psychological harm and post-traumatic disorders (Lousada, 2024). This supported the ruling issued in 2024 by Barcelona Social Court No. 28 on the classification as temporary disability of the condition of a content moderator employed by the company CCC Barcelona Digital Services. The worker presented panic attacks, avoidance behaviors, isolation at home, hypochondriacal rumination, a sensation of dysphagia, nocturnal awakenings and severe thanatophobia. The company's own website carried content stating that it was aware of the consequences of the job and announcing that its priority was to protect the mental health and well-being of its workers, for which purpose it made technology available to its employees and implemented AI-assisted moderation (Lousada, 2024). In 2022, in a landmark acknowledgement, Meta agreed to pay 52 million dollars to current and former moderators as compensation for the mental health problems they had developed in the course of their work (Newton, 2020).

The need to mitigate the effects on workers' mental health has driven the creation of algorithms that, in principle, facilitate the task of reviewing and classifying material on social media, using user engagement data (Wang et al., 2022) or neural networks and artificial intelligence systems (Agarwal et al., 2020; Alharthi et al., 2021; Andročec, 2020). These new moderation mechanisms promised to be a panacea and a solution to the serious crisis of credibility the sector was experiencing. During his testimony before Congress in 2018, Mark Zuckerberg highlighted dozens of times the role of AI as the future solution to content moderation (Gorwa et al., 2020).

In March 2020, as populations went into lockdown in response to the COVID-19 health and social crisis, Twitter and many other companies were forced to rely exclusively on automated moderation. The system showed serious shortcomings and, following this failure, many platforms apologized for the increase in errors (Gillespie, 2020). The debate on AI-based content moderation has both detractors and defenders. On this issue, the IFLA Statement on Libraries and Artificial Intelligence (Committee on Freedom of Access to Information, 2020) concludes that the use of AI in content moderation constitutes an infringement of freedom of expression and may lead to a lack of transparency and opacity regarding removal criteria. One of the most representative cases of this risk came to light in 2020 with the leak of TikTok's content moderation guidelines, which instructed the AI to remove posts created by physically "unattractive" users, including people with disabilities (Torres Vargas, 2022).

The need to establish clear rules to guide both the regulation of content by AI algorithms and human workers has led platforms to publish guideline policies, which can be summarized and combined in the following points:

High-priority content: child sexual exploitation and abuse, graphic violence, hate speech, harassment or bullying, extremist or terrorist content, content related to suicide or self-harm, dangerous disinformation, content inciting violence or crime, drug-related content and hate speech.

Moderate-priority content: disinformation, content related to illegal drugs, animal abuse, gender-based violence, offensive language, impersonation and disturbing graphic content.

Low-priority content: spam and fraud, conspiracy theories, harmful jokes or satire, low-level violence or disturbing situations.

Workers may be assigned to one of these content categories or even to all of them; this information is difficult to access, given the confidentiality agreements they are required to sign. This confidentiality is one of the major problems, as there is little information on their working conditions, work dynamics and, in general, the policies used by the large platforms. These clauses allow organizations to maintain practices marked by a serious lack of transparency (Torres Vargas, 2022).

As an alternative to automated content moderation and to encourage a more inclusive approach, X expanded its initiative known as Birdwatch, turning it into Community Notes. This system allows the user community to evaluate and contextualize the quality of the content disseminated—whether text, image or video—through collaborative notes visible to other users on the platform (Chuai et al., 2024). The measure was first implemented in the United States in 2023 and reached Spain in 2024, without breaching the measures established by the DSA, since X continues to submit its transparency reports, which state: "Our content moderation systems combine automated and human review with a robust appeals system that allows our users to quickly raise potential moderation anomalies or errors" (X Corp, 2024, n.p.).

In 2025, some organizations went a step further in moderation. In January, Meta's CEO, Mark Zuckerberg, issued a statement on Facebook announcing the end of content moderators, replacing them with a moderation system based on the platform's own users, following a model similar to the one implemented by X in the United States. However, Meta will continue to moderate content related to drugs, terrorism and child exploitation (Zuckerberg, 2025). Alexios Mantzarlis, Director of the Security, Trust, and Safety Initiative at Cornell University, quoted by Limón (2025) in El País, notes that "the decision not to fact-check and to relax content moderation opens the door to an increase in hate speech".

The many forms of harm that the spread of harmful content and the lack of safety on social media can cause, at a time of rapid change and great uncertainty regarding the decisions of platforms and public bodies, make it necessary to analyze human resources management in content moderation from the perspective of the Digital Services Act.

3. Objectives

To evaluate the deployment of human resources in content moderation on very large online platforms (VLOPs) with more than 45 million users, from the perspective of the Digital Services Act.

This will enable us to determine whether social media platforms allocate an adequate number of moderators relative to their active users/languages in the European Union (EU).

4. Methodology

A quantitative methodology is employed, based on documentary analysis, a review of reports, data analysis and an evaluation of the human resources used in content moderation by VLOPs operating in the European Union.

The research is carried out in two phases:

Phase 1: Review of the transparency reports for the second four-month period of 2023 of the main digital platforms (VLOPs) active in the European Union. The aim is to analyze the human resources employed in content moderation on the basis of the first reports resulting from the Digital Services Act (European Commission, 2023).

Data source: first transparency report, in compliance with the DSA, covering the second four-month period of 2023. The reports of the main platforms are used: YouTube, Meta (Facebook, Instagram), Pinterest, LinkedIn, Snapchat and X.

The aim is to extract quantitative information on human resources management on each platform, paying particular attention to the coverage of the official languages of the European Union.

Phase 2: Analysis and synthesis of the data obtained from the reports. We compute descriptive statistics, proportion measures and balance indices. The aim is to integrate the quantitative results obtained in the previous phase to develop a comprehensive understanding of the use of human resources in content moderation on digital platforms.

5. Data Analysis

The quantitative data extracted from the large platforms' reports on the number of moderators per language were used to evaluate the moderator-to-user ratio, calculated as the proportion of moderators per user for each social network using the following formula:

R mu = \frac{M}{U}

This ratio (R mu) yields a standardized metric that facilitates comparison between the different platforms. The results are shown in Table 1:

Table 1. Moderator-to-user ratio (Rmu)

Social network	R mu[1]
X	11.81
META	4.83
TIKTOK	41.85
YouTube	4.15
Snapchat	17.49
Pinterest	0.97
LinkedIn	3.25

Source: Authors' own elaboration

To assess the relative comparison (Rmu), the consistency of the platforms' R ratios is evaluated, revealing notable differences. The overall mean stands at 10.27 moderators per million users, with extremely high variability in dispersion, indicating a non-homogeneous distribution ($\sigma = 322289.72$), and reflecting a large disparity in the allocation of moderators per user, with some platforms significantly exceeding the mean (CV=31347.49%).

The ratio by language shows highly uneven results (Table 2). English leads the average ($\bar{x}$=298.15), far ahead of any other language and with an extreme range (998.6), mainly owing to the extremely high values for TikTok and Snapchat. The best-balanced languages are Spanish and Portuguese, with high means and similar medians, showing more uniform coverage. Languages such as Lithuanian and Greek have low means and zero medians, indicating little investment in human resources.

Table 2. Means, medians and ranges by language

Language	Mean *R*	Median *R*	Range (*max–mın*)
German	8.21	3.62	37.60
Bulgarian	6.28	1.42	36.32
Czech	5.53	0.00	32.63
Croatian	5.58	0.00	22.22
Danish	6.01	1.80	32.31
Slovenian	15.12	0.00	90.00
Slovak	7.15	0.00	48.89
Estonian	3.34	0.00	15.00
Spanish	8.98	6.33	24.64
Finnish	4.93	2.83	23.67
French	6.73	2.61	27.23
Greek	3.12	0.00	16.27
Dutch	4.55	0.81	23.86
Hungarian	4.27	0.00	23.33

[1] Moderator ratio per million users

English	298.15	25.35	998.60
Italian	5.07	2.50	22.16
Latvian	10.97	1.00	61.11
Lithuanian	2.02	0.00	7.50
Polish	4.44	2.77	19.62
Portuguese	16.44	13.33	45.05
Romanian	4.21	1.81	22.57
Swedish	6.19	1.50	33.75

Source: Authors' own elaboration

To analyze the distribution of the ratio across the different languages of the European Union (Rmui), the proportion of moderators per user by language is calculated (Table 3).

According to this ratio, English stands out above the rest (Rmui=298.15), possibly owing to the extremely high values on TikTok. Portuguese (Rmui=16.44) and Slovenian (Rmui=15.12) show values reflecting moderate moderation levels. Lithuanian (Rmui=2.02), Greek (Rmui=3.12) and Estonian (Rmui=3.34) are among the languages with the lowest ratios.

TikTok ($\sigma = 207.5$) and Snapchat($\sigma = 124.8$) show great variability in their values (Rmui) across languages, indicating that, although they maintain good averages, their coverage is not uniform, with some languages showing very high values and others very low ones. X also shows high variability ($\sigma = 97.06$), with very high values for English and values of 0 for a large proportion of languages.

META($\sigma = 5.04$) , Pinterest (4.57) σ =and LinkedIn ($\sigma = 5.50$)show less dispersion in the ratio, indicating more uniform coverage, although possibly limited in some languages.

The most striking deviation occurs on TikTok, which has a very high standard deviation, reflecting an intensive focus on some languages but a lack of coverage in others.

Table 3. Moderator-to-user ratio (Rmui) by language

Network	**R_{mu} /networks[2]**						
	X	META	TIKTOK	YouTube	Snapchat	Pinterest	LinkedIn
German	3.76	6.47	37.78	2.54	3.62	0.18	3.12
Bulgarian	1.42	4.55	36.32	1.67	0	0	0
Czech	0	2.97	32.63	3.13	0	0	0
Croatian	0	11.36	22.22	5.45	0	0	0
Danish	0	3.95	32.31	1.8	4.02	0	0
Slovenian	0	7.5	90	15	8.33	0	0
Slovak	0	0	48.89	1.16	0	0	0
Estonian	0	3	15	5.38	0	0	0
Spanish	0.93	4.72	25.57	10.02	13.92	1.40	6.33
Finnish	0	4.69	23.67	2.83	3.35	0	0
French	2.61	4.12	27.59	2.13	7.68	0.36	2.61
Greek	0	2.97	16.27	2.59	0	0	0
Dutch	0	5.05	23.86	0.81	0	0.2	1.9

[2] 0: platforms with 0 moderators

Hungarian	0	3.48	23.33	3.09	0	0	0
English	46	25.95	101.7	2.46	587.20	18.27	0
Italian	0.12	5.03	22.28	1.68	3.69	0.2	2.5
Latvian	0	1.82	12.86	61.11	0	1.1	0
Lithuanian	0	2.86	7.5	3.79	0	0	0
Polish	0	2.77	19.62	2.99	0.69	0	5
Portuguese	2.02	6.58	22.73	45.05	0	13	25
Romanian	0	2.87	22.57	2.21	1.81	0	0
Swedish	0	2.96	33.75	1.5	4.56	0.5	0

Source: Authors' own elaboration

The comparison of coverage by social network yields highly relevant data (Table 4). TikTok and YouTube have practically complete coverage, with both networks covering 22 languages with at least one moderator. Meta stands out with 95.45% coverage, lacking coverage in only one of the 24 official languages of the European Union, bearing in mind that Maltese is not covered in any case and Irish is assumed to fall under English. In the case of Snapchat (45.45%), Pinterest (40.91%) and LinkedIn (31.82%), coverage is more limited, with no moderators for half of the languages. Finally, X (31.82%) has low coverage, leaving a large number of languages without moderators.

Table 4. Comparison of coverage by social network

Social network	Languages covered (R>0)	Languages not covered (R=0)	Coverage (%)
X	7	15	31.82%
META	21	1	95.45%
TikTok	22	0	100.00%
YouTube	22	0	100.00%
Snapchat	10	12	45.45%
Pinterest	9	13	40.91%
LinkedIn	7	15	31.82%

Source: Authors' own elaboration

The relative performance of moderators gives us an idea of how moderator resources are distributed across networks and users (Table 5).

TikTok thus has the highest relative performance (72.52), combining high language coverage with a large investment in moderators, which suggests a well-distributed approach. X (66.66) and Snapchat (63.05) also show high values, albeit with lower language coverage, which indicates a more limited approach. The lowest values are found for Meta (5.48), YouTube (7.81), Pinterest (3.88) and LinkedIn (6.69), with a more uniform distribution but less intensity per language.

Table 5. Relative performance of moderators

Social network	Total moderators (ΣR)	Relative performance (moderators per language with R>0)
X	466.59	66.66
META	115.07	5.48
TikTok	1595.35	72.52
YouTube	171.72	7.81

Snapchat	630.54	63.05
Pinterest	34.92	3.88
LinkedIn	46.85	6.69

Source: Authors' own elaboration

This variability shows that some languages may depend heavily on a single network. To verify this, we calculated the percentage of each language's total that corresponds to each social network, which allows us to determine dependence (Table 6). According to this index, TikTok is the network on which most languages are most dependent, with dependence above 50%. YouTube stands out for Latvian (79.68%) and Portuguese (39.14%).

Languages such as Slovak and Slovenian depend almost exclusively on TikTok, which could represent a limitation for these countries.

Table 6. Dependence index by language

Language	Network with greatest dependence	Dependence (%)
German	TikTok	65.74%
Bulgarian	TikTok	82.62%
Czech	TikTok	84.25%
Croatian	TikTok	56.93%
Danish	TikTok	76.78%
Slovenian	TikTok	85.04%
Slovak	TikTok	97.68%
Estonian	TikTok	64.16%
Spanish	TikTok	40.66%
Finnish	TikTok	68.53%
French	TikTok	58.58%
Greek	TikTok	74.53%
Dutch	TikTok	74.89%
Hungarian	TikTok	78.03%
English	TikTok	47.85%
Italian	TikTok	62.74%
Latvian	YouTube	79.58%
Lithuanian	TikTok	53.00%
Polish	TikTok	63.15%
Portuguese	YouTube	39.14%
Romanian	TikTok	76.61%
Swedish	TikTok	77.94%

Source: Authors' own elaboration

6. Discussion

The analysis of the data presented reveals inequality across social networks with regard to the management of the human resources employed in content moderation. There are notable differences in the resources allocated by the platforms, which could affect the fairness and efficiency of moderation as required by the DSA. TikTok remains the leader in terms of mean and relative performance, despite not fully covering the official languages. Meta, YouTube and LinkedIn maintain a low level of resource allocation. Of all the networks analyzed, Pinterest has a ratio that could point to a shortfall in resource allocation. This precariousness could cause not only problems in content regulation, fostering more violent messages or less safe networks, but could also be affecting the workload of content moderators. In recent months, there has been a proliferation of

news reports on the psychosocial risks faced by this professional profile, due not only to work overload but also to constant exposure to shocking and emotionally exhausting images—all the more so if moderators must meet the minimum performance targets set by the companies themselves, which stand at around 95% correct identification of harmful content (Dang et al., 2018).

To analyze the adequacy of human resources in content moderation in detail, it would be essential to know the volume of content produced on each platform, but this figure does not appear in the transparency reports, nor is it a requirement under the DSA. Reporting the number of users is required, even though this figure may be of little use when analyzing resources, since a distinction must be made between active users, passive users and content-creating users. A network may have a large number of users who are mostly passive, producing a smaller volume of content than a network with fewer users but greater content-creation activity. Each platform's production volume should be taken into account in transparency reports, not only to certify safety on the network but also to safeguard the psychosocial health of moderators. For example, if we take the data on the volume of video uploaded to YouTube (Ceci, 2023)—720,000 hours of content per day—and establish the ratio of hours that each moderator would need to review, regardless of language, we would obtain a total of 389.8 hours per moderator, a ratio that is wholly beyond reasonable limits.

Another vitally important element in analyzing the adequate management of resources would be to examine the distribution of moderators by language. The results of this analysis reveal a clear predominance of moderation in English, which far exceeds that of the other languages, reflecting a significant concentration of moderator resources in this language. Languages such as Portuguese and Slovenian stand out with moderate values, implying acceptable coverage on some networks but limited coverage on others. Lithuanian, Greek and Estonian have significantly low figures, and some networks even have no moderators at all for these languages.

Networks such as Snapchat, Pinterest, LinkedIn and X have very limited coverage, with fewer than half of the languages represented, which suggests insufficient investment in multilingual moderation.

The networks with the greatest language coverage are TikTok and YouTube. In the case of TikTok, however, there is great inequality in the intensity of moderation by language, which could be due to a regional or commercial strategy that prioritizes certain language markets. These data also show an extreme dependence on TikTok, since for 20 of the 22 languages represented it accounts for more than 50% of all assigned moderators. Languages such as Slovak and Slovenian face significant risks should this network change its moderation policy.

In the case of YouTube, its role in Latvian and Portuguese stands out, which could represent an opportunity to diversify coverage in other languages.

As regards the balance of coverage by language, Spanish and Portuguese show the most balanced values, with high means and similar medians, reflecting more uniform moderation compared with other languages. The means for Lithuanian and Greek are the most

worrying, as they show low means with medians close to 0, which implies a lack of investment and an increased risk of harmful content being published in these languages.

These data reveal an imbalance in the distribution of moderators across the different languages of the European Union, from which Maltese—an official language of the EU—is completely excluded from the outset. The unequal distribution of resources across networks raises questions about fairness in moderation, as some language communities with low averages may be more exposed to harmful content owing to a lack of oversight. Moreover, the high dependence of certain languages on specific networks constitutes a high systemic risk, since a change in moderation policies could leave entire communities without adequate moderation.

7. Conclusions

The European Commission issued its assessment of the reports analyzed in this article, submitted for the second four-month period of 2023 in relation to compliance with the DSA, reporting only one complaint, against X, for lack of data transparency and failure to verify accounts. In addition, an investigation was opened into Meta for excessive misleading advertising, and another into TikTok concerning the protection of minors, owing to the addictive design of its TikTok Lite app, which would encourage minors to use the tool in exchange for rewards (Wittek, 2024).

In light of the Commission's assessment, we understand that the platforms' role with regard to the management of content moderation resources is considered entirely correct, despite the imbalance revealed in this study and without regard to the possible implications for the different communities.

Investment in human resources for content moderation must not only be substantial but also strategically distributed to meet users' specific needs in terms of language and culture.

The European Union has 24 official languages, and although the Digital Services Act does not explicitly state the need for content moderation to cover all official languages in full, the EU does support multilingualism and recognizes the need to acknowledge and protect languages, which enable greater competitiveness and social cohesion, as upheld in the Charter of Fundamental Rights of the EU (2000).

An equitable approach would mean that the number of moderators is adjusted in proportion to the number of users, the volume of content and linguistic diversity. Ensuring linguistic representation is essential to protect citizens' fundamental rights and guarantee an inclusive experience on the platforms. The exclusion of a language could lead to systematic lack of protection for its speakers, leaving a critical gap in the capacity to handle complaints or regulate harmful content. This situation not only puts platform users at risk but may also pose a threat to social cohesion and constitute a lack of respect for cultural diversity. It would be essential to analyze the ethical and social implications of unequal moderator allocation, especially in the context of minority languages.

An inadequate moderator-to-user ratio may also have serious consequences for the removal of harmful material, which may be delayed or even go unnoticed, resulting in a lack of control that creates a more hostile digital environment. The figures provided in the transparency reports in compliance with the DSA are insufficient to assess whether this ratio is adequate and whether technology companies are genuinely taking responsibility for investing fairly and strategically in content moderation.

These figures are also inadequate for analyzing the actual workload of each moderator and the impact of the nature of the content, information that would be essential to assess the real situation. Working under inadequate conditions may entail psychosocial risks that are not recognized as occupational diseases under the official schedule of occupational diseases, leaving workers unprotected.

The new law should require structural standards for the allocation of human and AI resources that protect workers and enable an adequate allocation of moderation resources, reflected in network safety while also respecting multiculturalism.

AI moderation requires human oversight, as shown by the results of the research conducted by the European Union Agency for Fundamental Rights in November 2023. The study analyzed 350,000 posts and found that, of 1,500 posts previously approved by content moderation (AI) tools, 53% were considered hate speech by human observers. When presenting the study, the Agency's director called for stronger content moderation to curb hate on social media:

"The sheer volume of hate we identified on social media clearly shows that the EU, its Member States and online platforms can step up their efforts to create a safer online space for all

that respects human rights, including freedom of expression. It is unacceptable to attack people online simply because of their gender, skin color or religion" (Vilas, 2023).

In the face of these conclusions, and precisely in the name of freedom of expression, platforms are abandoning content moderation, which has raised alarm in various forums, such as the World Economic Forum held in Davos in January 2025, which concluded that social networks' abandonment of content moderation will lead to a further erosion of democracy.

Under the banner of freedom of expression, the debate has drifted in dangerous directions that may cause not only this erosion of democracy but also serious harm at the individual level and at the level of society as a whole. Regulating content moderation is essential to guarantee safe communications. Large technology companies must invest in human resources and guarantee adequate working conditions that allow moderators to do their work without damaging their psychosocial integrity. Likewise, it is essential for the authorities to establish legislation to regulate social media content, as this helps to prevent disinformation, protect users' rights and guarantee a safer and more equitable digital environment for all.

References

Agarwal, A., Mittal, M., Pathak, A., & Goyal, L. M. (2020). Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning. *SN Computer Science*, *1*(3), 143. https://doi.org/10.1007/s42979-020-00165-4

Alharthi, R., Alhothali, A., & Moria, K. (2021). A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter. *Information Systems*, *99*, 101740. https://doi.org/10.1016/j.is.2021.101740

Andročec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, *12*(2), 205-216. https://doi.org/10.2478/ausi-2020-0012

Boyd, D & Marwick, A. (2017, January 4). *Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies*. Paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society". Microsoft Research. https://doi.org/10.31219/osf.io/2gec4

Ceci L. (2023, April). Statista. *Horas de video subidas a YouTube*. https://es.statista.com/estadisticas/599096/horas-de-video-subido-a-youtube-cada-minuto-2007/

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement with Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-52. https://doi.org/10.1145/3686967

Committee on Freedom of Access to Information (2020). *IFLA Statement on Libraries and Artificial Intelligence*. International Federation of Library Associations and Institutions (IFLA). https://repository.ifla.org/handle/20.500.14598/1646

European Commission (2023) *Paquete de Servicios Digitales* . https://digital-strategy.ec.europa.eu/en/factpages/safer-fairer-online-environment

Dang B., Riedl J., & Lease M. (2018). But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. *arXiv:1804.10999v4 [cs.HC] 5 Jan 2020*. https://doi.org/10.48550/arXiv.1804.10999

Évole Ramón, J., & Lara, R (Producers). (2024, October 13*) Redes Sociales: la fábrica del terror* (Television program). In Salvados. Atresplayer https://www.atresplayer.com/lasexta/programas/salvados/temporada-21/redes-sociales-la-fabrica-del-terror-parte-2_670949534911b0e4979c8b87/

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Grimmelmann, J. (2017). The Virtues of Moderation. *Yale Journal of Law & Technology*, 42-108. https://doi.org/10.31228/osf.io/qwxf5

Hernández García, V. (2021). Retos de traducción de las «Terms and Conditions» de las redes sociales: análisis jurídico y terminológico contrastivo inglés-español basado en corpus. *Hikma*, *20*(1), 125-156. https://doi.org/10.21071/hikma.v20i1.12938

Jiménez González, A., & Cancela Rodríguez, E. (2023). ¿Es posible gobernar a las plataformas digitales? Análisis crítico de la Ley Europea de Servicios Digitales. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, *20*(1), 91-99. https://doi.org/10.5209/tekn.82074

Lievens, E., Livingstone, S., McLaughlin, S., O'Neill, B., & Verdoodt, V. (2019). Children's Rights and Digital Technologies. In U. Kilkelly & T. Liefaard (Eds.), *International Human Rights* (pp. 487-513). Springer. https://doi.org/10.1007/978-981-10-4184-6_16

Limón, R. (2025, January 10). Suprimir la verificación y moderación en redes aumenta el odio y el acoso, advierten los expertos | Tecnología | EL PAÍS. *El País*. https://elpais.com/tecnologia/2025-01-10/suprimir-la-verificacion-y-moderacion-en-redes-aumenta-el-odio-y-el-acoso-advierten-los-expertos.html

Livingstone, S., & Third, A. (2017). Children and young people's rights in the digital age: An emerging agenda. *New Media & Society*, *19(5),* 657-670. https://doi.org/10.1177/1461444816686318

Lousada Arochena, J. F. (2024). Enfermedades del trabajo en la era digital: ¿están las leyes ajustadas a las nuevas realidades? Los trastornos psiquiátricos de un moderador de contenidos de Internet. *Revista de Jurisprudencia Labora*l, No. 2/2024. 1-6

Manovich, L. (2020). *Cultural analytics*. The MIT Press.

Mitchell, W. J. T. (2017). Iconology, visual culture and media aesthetics. *Poznańskie Studia Polonistyczne. Seria Literacka*, *30*, 341-364. https://doi.org/10.14746/pspsl.2017.30.17

Newton, C. (2020, May 12). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job - The Verge. *www.theverge.com*. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Nissenbaum, N. (2011). Privacy in Context: Technology, Policy, and Social Life. *German Law Journal. Vol 12.* 957-967 . https://doi:10.1017/S2071832200017168

Núñez-Cansado, M., López-López, A., & Somarriba-Arechavala, N. (2021). Covert advertising by kidsfluencers: A methodological proposal applied to the case study of the ten youngest youtubers with most followers in Spain. *Profesional de la Informacion*, *30*(2). https://doi.org/10.3145/epi.2021.mar.19

Roberts, S. (2019). *Behind the Screen*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Torres Vargas, G. Araceli. (2022). *Desafíos en el entorno de la información y la documentación ante las problemáticas sociales actuales*. Universidad Nacional Autónoma de México.

European Union. (2000). *Carta de los derechos fundamentales de la unión europea*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32000X1219%2801%29

European Union. (2022). *Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo, relativo a los servicios digitales (Ley de Servicios Digitales - DSA) .* Diario Oficial de la Unión Europea.https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32022R2065

Vilas C. (2023, November) *Un informe europeo pide reforzar la moderación de contenidos para frenar el odio en redes sociales. Euronews*. https://www.dailymotion.com/video/x8q2cuy

Wang, K., Fu, Z., Zhou, L., & Zhu, Y. (2022). Content Moderation in Social Media: The Characteristics, Degree, and Efficiency of User Engagement. *2022 3rd Asia Symposium on Signal Processing (ASSP)*, 86-91. https://doi.org/10.1109/ASSP57481.2022.00022

Wittek R. (2024, July).¿Cómo se ha aplicado la Ley de Servicios Digitales durante su primer año en vigor? *Euronews*. https://es.euronews.com/next/2024/07/22/como-se-ha-aplicado-la-ley-de-servicios-digitales-durante-su-primer-ano-en-vigor

X Corp. (2024). *DSA Transparency Report - October 2024*. https://transparency.x.com/dsa-transparency-report.html

Zuckerberg, M. (2025). *It's time to get back to our roots around free expression [Video]*. https://www.facebook.com/watch/?v=1525382954801931