VISUAL REVIEW | Vol. 18, No. 1, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.6128

HUMAN RESOURCE MANAGEMENT IN SOCIAL MEDIA CONTENT MODERATION: Perspectives from the Digital Services Act

MARIAN NÚÑEZ-CANSADO[1]
mariangeles.nunez@uva.es
ÁNGEL QUINTANA GÓMEZ [2]
angel.quintana@ulpgc.es
[1] Universidad de Valladolid, España
[2] Universidad de las Palmas de Gran Canarias, España

KEYWORDS

Content Moderation
Social Media
Digital Services Act (DSA)
Communication Ethics
European Union (EU)
Communication
Advertising

ABSTRACT

This study examines the implementation of human resources for content moderation on large digital platforms (VLOPs) within the European Union, employing a quantitative methodology based on documentary and data analysis. The results reveal substantial disparities in the ratio of moderators per language, with TikTok and YouTube demonstrating the most comprehensive coverage, while LinkedIn and X exhibit significant gaps. The study concludes that, although these platforms comply with regulatory requirements, incomplete language coverage and substandard working conditions pose considerable risks to both vulnerable communities and the moderators themselves.

Received: 13/ 07 /2025
Accepted: 28/ 10 / 2025

1. Introduction

The ability to store and share content online on a large scale has significantly transformed interpersonal interactions, conferring on social media a crucial role in personal communication, identity construction, and cultural expression (Manovich, 2020; Mitchell, 2017). However, this proliferation of content presents considerable ethical and privacy dilemmas, particularly regarding the management of digital images, encompassing their selection, organisation, and display, as well as their potentially violent or hate-inciting nature.

Content editing on digital platforms raises issues related to copyright, the consent of individuals depicted, and the handling of personal data involved in the published material, among other concerns. Although social media enables rapid and widespread content dissemination, platform terms and conditions often grant the service providers extensive property rights over user-generated content. As Hernández García (2021) notes, this contractual mechanism implies that users transfer rights of use to the platform, permitting reuse within the boundaries defined in the user agreement. Extensive literature highlights the problem of unauthorised content use and the violation of individual privacy, which can compromise the security of network users (Nissenbaum, 2011). Furthermore, the decontextualised reuse of content on digital platforms can result in plagiarism or distortions of representation, negatively influencing public perception of certain individuals or social groups (Boyd & Marwick, 2017). Additionally, content that promotes violence, incites hatred, or exploits children can profoundly affect both individual and collective identity formation.

These concerns are particularly acute in the case of children, a group highly vulnerable to the consequences of digital content dissemination (Lievens et al., 2019; Livingstone & Third, 2017). Risks include digital exploitation of images and identity, which can foster a distorted self-concept and lead to serious psychosocial problems (Núñez-Cansado et al., 2021).

Alongside the multiple risks associated with social media content and privacy, the role of platform algorithms must be considered. These systems, designed to maximise user engagement and time spent online, tend to prioritise sensationalist, polarising, or emotionally charged content. In 2021, a former Facebook-Meta employee raised serious concerns regarding the platform's management by leaking numerous documents to *The Wall Street Journal*. These documents revealed practices that negatively affected younger users, derived from both the content and the algorithms employed. In testimony before the European Parliament, the employee stated: '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 economic resources' (Jiménez González & Cancela Rodríguez, 2023, p. 91).

Arturo Béjar, head of Facebook's Integrity and Care department, confirmed in the documentary *Social Media: The Terror Factory* that the platform was aware of the harm being caused but failed to act. In June 2020, Béjar sent an email to Mark Zuckerberg and senior management presenting alarming data, such as 51% of Instagram users reporting negative content and 21.8% of teenagers aged 13 to 15 experiencing direct harassment, asserting that Instagram was likely responsible for the greatest incidence of sexual harassment in human history (Évole Requena & Lara, 2024).

This evidence has prompted a new debate regarding the dangers of social media and highlighted the necessity of implementing regulatory systems to monitor platform content. Although legal provisions governing data use and protection, notably the European General Data Protection Regulation (GDPR), already existed, these were deemed insufficient in addressing the reported harms. Consequently, in 2017 Germany enacted the Network Enforcement Act (NetzDG), which formed the basis for various initiatives, including the European Code of Practice on Disinformation (2018) and the Spanish National Security Council's ministerial provision of 30 October 2020 establishing procedures for action against disinformation. UNESCO has also engaged with the issue, organising seminars on content moderation and freedom of expression on social media, covering moderation criteria, AI versus human intervention, transparency, and platform accountability.

The absence of specific regulation and the challenges of moderating digital content prompted cooperation between corporations and governments, culminating in the European Union's enactment of the Digital Services Act (DSA), which came into force on 25 August 2023 (European Union, 2022). The DSA represents a pivotal legislative measure for ensuring a safe digital environment, introducing obligations such as algorithmic transparency, prohibition of misleading and intrusive advertising, enhanced reporting mechanisms, and the removal of illegal content. It also establishes protective

measures for users, particularly minors. Platforms categorised as very large online platforms (VLOPs) with over 45 million active EU users are required to implement additional measures to address systemic risks and report regularly on content moderation and political advertising. The primary objective of the DSA is to balance security, transparency, and responsibility within the EU's digital ecosystem, holding platforms accountable for hosted content while safeguarding users from harmful material, abusive algorithms, and misuse of personal data.

Nevertheless, these laws face significant implementation and enforcement challenges within a context of continual digital transformation, where content moderation remains complex due to conflicts with commercial interests or the absence of specific regulations in certain virtual spaces (Jiménez González & Cancela Rodríguez, 2023).

2. Content Moderators

The risks arising from the misuse of content on social media have intensified the need not only for comprehensive legislative analysis, but also for active commitment from social media corporations. The diversity of content, recommendation algorithms, and social graphs has generated numerous complaints and criticisms directed at content platforms, obliging corporations to maintain constant and rigorous monitoring. This responsibility poses both logistical and public relations challenges (Gillespie, 2020). Consequently, the practice of 'commercial content moderation' or 'platform moderation' has become increasingly significant. In recent years, the scholarly literature has expanded to examine key issues surrounding this emergent professional profile, including employment conditions, legislative frameworks, democratic legitimacy, and the lack of transparency and accountability within large corporations (Gorwa et al., 2020).

Content moderation is defined as the organised practice of filtering user-generated content on digital platforms (Roberts, 2019) and, according to Grimmelmann (2017, p. 18), as 'governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse'. Its purpose is to ensure a safe and respectful environment for all users, preventing the dissemination of inappropriate, illegal, or harmful content, such as hate speech, violence, misinformation, harassment, or explicit material. Moderation may occur proactively, prior to content dissemination, or reactively, in response to complaints from users, site administrators, or affected entities.

Over the past year, the professional profile of content moderators has sparked debate around three fundamental aspects: the development of standardised manuals for professional practice, recognition of psychosocial risks and their classification as occupational diseases, and the delineation of boundaries between human moderation and AI-based moderation.

In 2022, the European Agency for Safety and Health at Work (OSHA) published Occupational Safety and Health Risks of Online Content Review Work Provided through Digital Labour Platforms, identifying exposure to violent, criminal, abusive, or illegal content as a source of stress capable of inducing psychological harm and post-traumatic disorders (Lousada Arochena, 2024). This assessment supported the decision of Barcelona Social Court No. 28 (2024), which classified a content moderator employed by CCC Barcelona Digital Services as temporarily disabled due to work-related psychological trauma. The moderator experienced panic attacks, avoidance behaviour, home isolation, hypochondriacal rumination, dysphagia, nocturnal awakenings, and significant thanatophobia. The company publicly acknowledged the mental health risks inherent in content moderation and committed to safeguarding employee well-being through technological tools and AI-assisted moderation (Lousada, 2024).

In a landmark case in 2022, Meta agreed to pay $52 million in compensation to current and former moderators suffering mental health problems related to their work (Newton, 2020).

To mitigate such risks, platforms have developed algorithms intended to facilitate content review and classification, leveraging user engagement data (Wang et al., 2022) and employing neural networks and artificial intelligence systems (Agarwal et al., 2020; Alharthi et al., 2021; Andročec, 2020). These technological interventions were initially hailed as a solution to the sector's crisis of credibility. Mark Zuckerberg repeatedly emphasised AI as a key future solution for content moderation in his 2018 Congressional testimony (Gorwa et al., 2020).

During the COVID-19 lockdown in March 2020, Twitter and other platforms were compelled to rely exclusively on automated moderation. The systems exhibited significant shortcomings, leading to

increased errors and subsequent public apologies from several platforms (Gillespie, 2020). The debate surrounding AI-based content moderation continues, with both proponents and critics. The IFLA Statement on Libraries and Artificial Intelligence (Committee on Freedom of Access to Information, 2020) asserts that AI in content moderation may violate freedom of expression and engender opacity regarding removal criteria. A prominent example of this risk emerged in 2020, when leaked TikTok content moderation guidelines revealed AI directives to remove posts by users deemed physically 'unattractive', including individuals with disabilities (Torres Vargas, 2022).

The necessity of establishing clear rules to guide both AI-driven and human content moderation has led platforms to publish policy guidelines, which can be summarised and grouped as follows:

- High-priority content: child sexual exploitation and abuse, graphic violence, hate speech, harassment or bullying, extremist or terrorist content, content relating to suicide or self-harm, dangerous misinformation, content inciting violence or crime, and drug-related content.
- Moderate-priority content: misinformation, content related to illegal drugs, animal abuse, gender-based violence, offensive language, impersonation, and disturbing graphic material.
- Low-priority content: spam and fraud, conspiracy theories, harmful joke or satirical content, low-level violence, or disturbing situations.

Workers may be assigned to one or more of these content categories; however, access to this information is often restricted due to the confidentiality agreements they are required to sign. Such confidentiality represents a significant barrier to transparency, as there is limited information regarding their working conditions, work dynamics, and, more broadly, the internal policies of large platforms. These clauses facilitate organisational practices characterised by considerable opacity (Torres Vargas, 2022).

As an alternative to fully automated content moderation and to foster a more participatory approach, X expanded its initiative previously known as Birdwatch, transforming it into *Community Notes.* This system enables the user community to evaluate and contextualise the quality of disseminated content, whether text, image, or video, through collaborative notes visible to other users on the platform (Chuai et al., 2024). Initially implemented in the United States in 2023, the system was introduced in Spain in 2024 without contravening the Digital Services Act (DSA), as the platform continues to submit transparency reports stating: 'Our content moderation systems combine automated and human review with a robust appeal system that allows our users to quickly raise potential anomalies or moderation errors' (X Corp, 2024, s.p.).

In 2025, some organisations have gone further in restructuring moderation practices. In January, Meta CEO Mark Zuckerberg announced via Facebook that the platform would phase out traditional content moderators, replacing them with a moderation system relying primarily on platform users, following a model similar to that implemented by X in the United States. The platform, however, will continue to moderate content relating to drugs, terrorism, and child exploitation (Zuckerberg, 2025). Alexios Mantzarlis, Director of Security, Trust, and Safety at Cornell University, cited in El País by Limón (2025), warned that 'the decision not to verify data and to relax content moderation opens the door to an increase in hate speech'.

Given the multiple harms associated with the dissemination of unsafe content, coupled with the ongoing uncertainty surrounding the decisions of platforms and regulatory authorities, it is essential to analyse human resource management in content moderation from the perspective of the Digital Services Act.

3. Objectives

Assess the implementation 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 analysis enables an evaluation of whether social media platforms allocate an adequate number of moderators relative to their active users and languages within the European Union (EU).

4. Methodology

A quantitative methodology was employed, based on documentary analysis, review of reports, data analysis, and evaluation of the human resources used in content moderation on VLOPs actively operating within the European Union.

The research was conducted in two phases:

Phase 1: Review of transparency reports for the second quarter of 2023 from the main VLOPs operating in the European Union (European Commission, 2023). The objective was to analyse the human resources employed in content moderation, based on the first transparency reports produced under the Digital Services Act.

Data source: First transparency reports, in compliance with the DSA, corresponding to the second quarter of 2023. Reports from the principal platforms were used: YouTube, Meta (Facebook and Instagram), Pinterest, LinkedIn, Snapchat, and X. The aim was to extract quantitative information on human resource management, with 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. Descriptive statistics, proportion measures, and balance indices were calculated. This phase aimed to integrate the quantitative results from the previous stage to develop a comprehensive understanding of the deployment of human resources in content moderation across digital platforms.

5. Data Analysis

The quantitative data extracted from the reports of the major platforms, based on the number of moderators per language, were evaluated using the moderator-to-user ratio, calculated as the proportion of moderators per user on each social network according to the following formula:

Rmu=MUR_{mu} = \frac{M}{U}Rmu=UM

This ratio (RmuR_{mu}Rmu) provides a standardised metric, enabling comparison across different platforms. The results are presented in Table 1:

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

Social Network	R mu[1]
X	11.81
META	4.83
TIK TOK	41.85
YouTube	4.15
Snapchat	17.49
Pinterest	0.97
LinkedIn	3.25

Source: Author's elaboration, 2024.

To evaluate the relative comparison (RmuR_{mu}Rmu), the consistency of the R proportions across platforms was assessed, revealing notable differences. The overall average was 10.27 moderators per million users, accompanied by extremely high variability in dispersion, indicating a non-homogeneous distribution (σ=322,289.72\sigma = 322{,}289.72σ=322,289.72), and reflecting a considerable disparity in the allocation of moderators per user, with some platforms significantly exceeding the mean (CV = 31{,}347.49%).

The ratio by language demonstrates highly disparate results (Table 2), with English leading the average (x¯=298.15\bar{x} = 298.15x¯=298.15) far above any other language and displaying an extreme range (998.6), largely attributable to outlying values in TikTok and Snapchat. The most balanced languages are Spanish and Portuguese, which exhibit high averages and closely aligned medians, indicating more uniform coverage. Conversely, languages such as Lithuanian and Greek show low averages and zero medians, indicating minimal investment in human resources.

[1] Ratio of moderators per million users

Table 2. Averages, medians and ranges by language

Language	Average *R*	Median *R*	Range (*max–min*)
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
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: Author's elaboration, 2024.

To analyse the distribution of the moderator-to-user ratio across the different languages of the European Union (RmuiR_{mui}Rmui), the ratio of moderators to users by language was calculated (Table 3). According to this metric, English stands out above the rest (Rmui=298.15R_{mui} = 298.15Rmui=298.15), likely due to the extremely high values observed on the TikTok platform. Portuguese (Rmui=16.44R_{mui} = 16.44Rmui=16.44) and Slovenian (Rmui=15.12R_{mui} = 15.12Rmui=15.12) exhibit moderate levels of moderation, while Lithuanian (Rmui=2.02R_{mui} = 2.02Rmui=2.02), Greek (Rmui=3.12R_{mui} = 3.12Rmui=3.12), and Estonian (Rmui=3.34R_{mui} = 3.34Rmui=3.34) are among the languages with the lowest ratios.

The TikTok (σ=207.5\sigma = 207.5σ=207.5) and Snapchat (σ=124.8\sigma = 124.8σ=124.8) platforms display considerable variability in RmuiR_{mui}Rmui values across languages, indicating that, despite high averages, coverage is uneven, with some languages receiving extensive moderation and others very little. The X platform similarly demonstrates high variability (σ=97.06\sigma = 97.06σ=97.06), with very high values in English but zero coverage in most other languages.

Conversely, Meta (σ=5.04\sigma = 5.04σ=5.04), Pinterest (σ=4.57\sigma = 4.57σ=4.57), and LinkedIn (σ=5.50\sigma = 5.50σ=5.50) show less dispersion, indicating more uniform moderation across languages, although some languages may still be inadequately covered. TikTok exhibits the most pronounced deviation, with a very high standard deviation reflecting intensive focus on selected languages and a lack of coverage in others.

Table 3. Moderator-to-user ratio (Rmui)/languages

Network	R_{mu} /networks[2]						
	X	META	TIK TOK	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
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: Author's elaboration, 2024.

The comparison of coverage by social network reveals several notable findings (Table 4). TikTok and YouTube provide almost complete coverage, with both platforms offering at least one moderator for 22 languages. Meta demonstrates high coverage at 95.45%, lacking moderation in only one of the 24 official languages of the European Union, noting that Maltese is not covered in any case and that Irish is assumed to be included under English. In contrast, Snapchat (45.45%), Pinterest (40.91%), and LinkedIn (31.82%) display more limited coverage, with approximately half of the languages having no assigned moderators. Finally, the X platform exhibits low coverage (31.82%), leaving a substantial number of languages without any moderation.

Table 4. Comparison of coverage by social network

Social Network	Languages with Coverage (R>0)	Languages without Coverage (R=0)	Percentage of Coverage (%)
X	7	15	31.82
TARGET	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: Author's elaboration, 2024.

The relative performance of moderators provides insight into the allocation of moderation resources across networks and user populations (Table 5). TikTok demonstrates the highest relative performance (72.52), reflecting both extensive linguistic coverage and substantial investment in moderators, indicative of a well-distributed approach. The X platform (66.66) and Snapchat (63.05) also exhibit high values, albeit with more limited linguistic coverage, suggesting a narrower distribution

[2] 0: platforms with 0 moderators

strategy. Lower values are observed in Meta (5.48), YouTube (7.81), Pinterest (3.88), and LinkedIn (6.69), reflecting a more uniform allocation across languages but with lower 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
GOAL	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: Author's elaboration, 2024.

This variability indicates that certain languages may rely heavily on a single platform. To assess this, we calculated the total percentage of moderators assigned to each language per social network, providing a measure of dependency (Table 6). According to this index, TikTok exhibits the highest dependency for most languages, with values exceeding 50%. YouTube is particularly notable for Latvian (79.68%) and Portuguese (39.14%). Languages such as Slovak and Slovenian are almost entirely dependent on TikTok, which may present a significant limitation for these communities.

Table 6. Dependency index by language

Language	**Network with Highest Dependency**	**Percentage of Dependency (%)**
German	TikTok	65.74
Bulgarian	Tik Tok	82.62%
Czech	TikTok	84.25%
Croatian	TikTok	56.93
Danish	Tik Tok	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: Author's elaboration, 2024.

6. Discussion

The analysis of the data provides insight into inequalities in social media with respect to the management of human resources employed in content moderation. There are significant differences in the allocation of resources across platforms, which may compromise both the fairness and efficiency of

moderation, as mandated by the DSA. TikTok continues to lead in terms of average and relative performance, despite not achieving full coverage in official languages. Meta, YouTube, and LinkedIn display low levels of resource allocation, while Pinterest exhibits ratios indicative of potentially insufficient investment. This precariousness may not only impede effective content regulation, thereby permitting more violent or unsafe content, but may also adversely affect the workload and well-being of moderators. Recent reports highlight the psychosocial risks associated with this professional role, stemming not only from excessive workloads but also from continuous exposure to distressing and emotionally taxing material. Such risks are exacerbated when moderators are required to meet corporate performance standards, which often demand approximately 95 per cent accuracy in the identification of harmful content (Dang et al., 2018).

To comprehensively assess the adequacy of human resources in content moderation, information on the volume of content produced by each platform is essential. However, this data is neither included in transparency reports nor required under the DSA. While platforms are obliged to report user numbers, this measure is of limited utility for resource analysis, as a distinction must be drawn between active users, passive users, and content creators. A network with a large number of mostly passive users may generate considerably less content than a smaller network with higher content creation activity. The inclusion of production data in transparency reports would not only enhance network security but also safeguard the psychosocial health of moderators. For instance, considering YouTube's daily volume of 720,000 hours of content (Ceci, 2023) and calculating the workload per moderator without accounting for language differences, each moderator would theoretically need to evaluate 389.8 hours of content, which is entirely unfeasible.

Another piece of data of vital importance for analysing the adequate management of resources is the distribution of moderators by language. The results indicate a clear predominance of moderation in English, which far exceeds that of other languages and reflects a significant concentration of moderator resources in this language. Portuguese and Slovenian exhibit moderate values, indicating acceptable coverage in some networks but limited coverage in others. Lithuanian, Greek, and Estonian show notably low figures, and in some networks, there is a complete absence of moderators.

Networks such as Snapchat, Pinterest, LinkedIn, and X display very limited coverage, representing less than half of the languages, which suggests insufficient investment in multilingual moderation. The platforms with the greatest linguistic coverage are TikTok and YouTube. In the case of TikTok, however, there is substantial inequality in the intensity of moderation depending on the language, which may reflect a regional or commercial strategy prioritising certain linguistic markets. The data also indicate extreme dependence on TikTok, which accounts for more than 50% of all assigned moderators in 20 of the 22 languages considered. Languages such as Slovak and Slovenian are therefore exposed to significant risks if there is a change in the network's moderation policies.

YouTube plays a notable role in moderation for Latvian and Portuguese, potentially providing an opportunity to diversify coverage in other languages. Spanish and Portuguese show the most balanced coverage, with high averages and close medians, reflecting more uniform moderation compared to other languages. By contrast, Lithuanian and Greek display concerning averages with medians close to zero, indicating low investment and an increased risk of harmful content being disseminated in these languages.

These results reveal an imbalance in the distribution of moderators across the languages of the European Union, with Maltese completely excluded. The unequal allocation of resources across networks raises questions about fairness in moderation. Language communities with low coverage may be disproportionately exposed to harmful content due to insufficient supervision, while high dependence on a single network constitutes a systemic risk, as changes in moderation policies could leave entire communities without adequate content oversight.

7. Conclusions

The European Commission issued a statement regarding the reports analysed in this article, submitted in the second quarter of 2023 in relation to compliance with the Digital Services Act. The statement noted only one complaint against network X for lack of data transparency and insufficient account verification. In addition, an investigation was launched into the META network for excessive misleading advertising, and TikTok was investigated in relation to the protection of minors, specifically concerning

the addictive design of its TikTok Lite application, which could encourage minors to use the platform in exchange for rewards (Wittek, 2024).

In light of the Commission's ruling, the role of platforms in managing content moderation resources is considered compliant, despite the imbalances highlighted in this study and without taking into account the potential implications for different language communities. The allocation of human resources to content moderation must not only be substantial, but also strategically distributed to address the specific needs of users with regard to language and culture.

There are twenty-four official languages in the European Union, and although the Digital Services Act does not explicitly require coverage of all official languages in content moderation, the legislation supports multilingualism and recognises the need to protect and promote languages, thereby fostering competitiveness and social cohesion, in line with the Charter of Fundamental Rights of the European Union (2000).

An equitable approach would involve adjusting the number of moderators in proportion to user volume, content volume, and linguistic diversity. Ensuring linguistic representation is essential to protect the fundamental rights of citizens and to guarantee an inclusive user experience on digital platforms. The exclusion of a language could systematically disadvantage its speakers, creating a critical gap in the capacity to respond to complaints or to regulate harmful content. This situation not only places users at risk but also threatens social cohesion and undermines respect for cultural diversity. It is therefore crucial to analyse the ethical and social implications of inequality in the allocation of moderators, particularly in contexts involving minority languages.

An inadequate ratio of moderators to users can have serious consequences for the timely removal of harmful material, which may be delayed or go unnoticed, thereby creating a more hostile digital environment. The figures provided in transparency reports on compliance with the Digital Services Act are insufficient to determine whether this ratio is adequate or whether technology companies are genuinely investing in content moderation in a fair and strategic manner.

These data are also insufficient for a true assessment of the actual workload of individual moderators and the impact of the content they review, which is essential for evaluating the real working conditions. Performing such work under inadequate conditions poses psychosocial risks that are not recognised as occupational diseases under the current regulatory framework, leaving workers without adequate protection.

New legislation should establish structural standards for the allocation of both human and AI resources, ensuring worker protection while facilitating effective content moderation. Such measures would enhance network security and respect multiculturalism simultaneously.

AI-based moderation requires human oversight, as demonstrated by research conducted by the European Union Agency for Fundamental Rights in November 2023. The study analysed 350,000 posts and found that of 1,500 posts previously evaluated by approved AI moderation tools, 53% were classified as hate speech by human observers. In presenting these findings, the Agency’s director called for stronger content moderation to combat hate on social media stating, ‘The large volume of hate speech we identified on social media clearly shows that the EU, its Member States and digital platforms can step up their efforts to create a safer online space for everyone that respects human rights, including freedom of expression. It is unacceptable to attack people online simply because of their gender, skin colour, or religion’ (Vilas, 2023).

In light of these findings, and ironically in the name of freedom of expression, some platforms have begun to retreat from content moderation. This trend has raised concern in multiple forums, including the World Economic Forum in Davos in January 2025, where it was concluded that the refusal of social media platforms to moderate content risks further eroding democracy.

The debate framed around freedom of expression is taking a dangerous turn, with potential consequences not only for democratic governance but also for the well-being of individuals and society as a whole. Regulation of content moderation is therefore essential to ensure secure communication. Large technology companies must invest in human resources and guarantee adequate working conditions, enabling professional development without compromising the psychosocial integrity of moderators. At the same time, authorities must enact and enforce legislation governing content on social media, thereby mitigating misinformation, protecting users’ rights, and promoting 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, enero 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, abril). 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

Comisión europea (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 Requena J., & Lara, R (Productores). (2024, octubre 13*) Redes Sociales: la fábrica del terror* (Programa de televisión). En 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. En 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, enero 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, mayo 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.

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

Unión Europea. (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, noviembre) *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, julio).¿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