

A Year of the DSA Transparency Database: What it (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election

Gautam Kishore Shahi^{1*}, Benedetta Tessa^{2,3*}, Amaury Trujillo², Stefano Cresci²

¹University of Duisburg-Essen, Germany

²IIT-CNR, Italy

³University of Pisa, Italy

gautam.shahi@uni-due.de, benedetta.tessa@iit.cnr.it, amaury.trujillo@iit.cnr.it, stefano.cresci@iit.cnr.it

Abstract

Social media platforms face heightened risks during major political events. Yet, how platforms adapt their moderation practices in response remains unclear. The Digital Services Act Transparency Database offers an unprecedented opportunity to systematically study content moderation at scale, enabling researchers and policymakers to assess platforms' compliance and effectiveness. Herein, we analyze 1.58 billion self-reported moderation actions taken by eight large social media platforms during an extended period of eight months surrounding the 2024 European Parliament elections. Our findings reveal a lack of adaptation in moderation strategies, as platforms did not exhibit significant changes in their enforcement behaviors surrounding the elections. This raises concerns about whether platforms adapted their moderation practices at all, or if structural limitations of the database concealed possible adjustments. Moreover, we found that noted transparency and accountability issues persist nearly a year after initial concerns were raised. These results highlight the limitations of current self-regulatory approaches and underscore the need for stronger enforcement and data access mechanisms to ensure that online platforms uphold their responsibility in safeguarding democratic processes.

Introduction

The European Parliament elections are paramount in shaping European Union (EU) politics. Held every five years, they allow citizens to elect their representatives for EU's legislative power, with the newly elected Parliament in turn electing the President of the European Commission. The most recent European Parliament elections took place in 2024—a year with an extraordinary number of national elections worldwide—from June 6th to 9th across all EU countries. The President of the commission was later elected via a secret ballot on July 18th. Such national and supranational elections have a major impact on social media platforms, which play a crucial role in hosting political campaigns and disseminating election-related information (Rho and Mazmanian 2020; Papakyriakopoulos, Engelmann, and Winecoff 2023; Shahi et al. 2024). For example, politicians often engage with citizens and encourage political learning

as well as electoral participation through their social media channels (Cresci et al. 2014; Fatema, Yanbin, and Fugui 2022; Kim and Ellison 2022; Bene et al. 2022). At the same time however, online electoral discourse can also be targeted by information tampering actions orchestrated to gain pre-electoral consent (Chen, Chen, and Xia 2022). These activities include information manipulation (Cinelli et al. 2020; Tardelli et al. 2020; Matatov, Naaman, and Amir 2022; Mazza et al. 2022), the use of deepfakes (Haq et al. 2024; Diakopoulos and Johnson 2021), targeted harassment (Hua, Naaman, and Ristenpart 2020), as well as opaque and unfair political advertising (Bär et al. 2024).

These phenomena are examples of the risks associated with periods of increased online political activity, where critical interests—both political and economic—come into play. In response, social media platforms attempt to mitigate the issues through content moderation (Gillespie 2018). During electoral periods, online platforms may intensify efforts to ensure trustworthy online discourse through the enforcement of multiple content and account moderation actions (Pierri et al. 2023; Majó-Vázquez et al. 2021; Shahi and Mejova 2025; Cima et al. 2025).

Content moderation has recently gained increased relevance not only for platform users but also for European regulators. In October 2022, the European Union enacted the Digital Services Act (DSA) to regulate online platforms and foster a more transparent, inclusive, and safe digital environment (European Parliament and Council 2022). Among its requirements, the DSA obliges large online platforms to report all their moderation actions within the EU by submitting clear, detailed, and timely *statement of reasons* (SoRs) to the DSA Transparency Database (DSA-TDB)¹—an open and centralized repository hosted by the European Commission (Trujillo, Fagni, and Cresci 2025; Kaushal et al. 2024). Operational since September 2023, the DSA-TDB represents an unprecedented tool for transparency and promised to revolutionize the observability of online platforms. For this reason, a few early works have analyzed the initial information that platforms submitted to the database during its first months of operation (Trujillo, Fagni, and Cresci 2025; Dergacheva et al. 2023; Aspromonte et al. 2024; Kaushal et al. 2024; Drolsbach and Pröllochs 2024). However, these

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://transparency.dsa.ec.europa.eu/>

uncovered compliance deficiencies by platforms and highlighted significant issues in the structure of the database itself, all of which limit its overall usefulness and reliability (Trujillo, Fagni, and Cresci 2025; Kaushal et al. 2024).

Here, we carry out the most extensive analysis of the DSA-TDB to date, almost one year after its initial release. We consider a broad observation period of ten months surrounding the 2024 EU elections and we explore the self-reported moderation actions of the eight largest social media platforms in the EU. By analyzing 1.58B SoRs in a politically critical time, we seek to understand whether the DSA-TDB has lived up to its transparency promise, assessing if and how the reported moderation practices of large social media platforms changed in response to the heightened integrity risks. Additionally, our extensive exploratory analysis allows assessing whether the shortcomings reported in previous studies have been addressed. This work thus answers the following research questions:

- **RQ1:** *How did self-reported moderation practices change during the European electoral period?* We aim to identify possible notable shifts in content moderation practices before and after the 2024 EU elections, such as changes in the volume and type of moderated content.
- **RQ2:** *To what extent has the reliability and consistency of the database improved since its initial release?* Initial analyses identified significant issues that limited the practical utility of the database as a transparency tool. Our study revisits these concerns with a larger and recent dataset, assessing progress and persistent challenges in achieving meaningful platform transparency.

Our results contribute to advancing the understanding of transparency mechanisms in digital governance, inform future regulatory decisions, and provide a timely resource for policymakers, scholars, and platforms aiming to foster greater integrity and accountability in online spaces.

Related work

The introduction of the European DSA-TDB has marked a significant milestone in digital legislation worldwide, aiming to enhance fairness and transparency in online governance. Hence, despite being launched only in September 2023, it has already been the focus of multiple studies assessing its effectiveness and the compliance of involved platforms. For instance, an initial study analyzed a 10-day sample covering all platforms to evaluate whether the database met its transparency and fairness objectives (Kaushal et al. 2024). While this study acknowledged the value of the insights into moderation decisions, it also identified inconsistencies and non-compliance issues arising from the self-reported nature of the data. Other analyses further highlighted heterogeneity in content moderation practices, including differences in the types of moderated content, the application of visibility restrictions, and the reported use of automation (Drolsbach and Pröllochs 2024; Dergacheva et al. 2023). Extending these early investigations, a 100-day study found that while platforms formally comply with the requirements of the DSA-TDB, they often omit key optional details in their SoRs, limiting the database’s prac-

tical utility (Trujillo, Fagni, and Cresci 2025). This study also revealed substantial discrepancies in self-reported moderation practices across platforms, suggesting varying levels of adherence to the intended structure of the DSA-TDB. Moreover, cross-checking the database against platforms’ own transparency reports exposed significant inconsistencies in the submitted data. Beyond direct assessments of the database’s quality, other work has explored ways to enhance its usability. For instance, Aspromonte et al. (2024) employed a multi-agent system based on large language models (LLMs) to link SoRs to the corresponding sections of platforms’ Terms of Service. Their findings suggest that LLMs can provide valuable contextualization, improving user understanding of moderation decisions and potentially increasing engagement with the DSA.

While these studies offer critical insights, they were conducted within the first months after the DSA-TDB’s release. Now—more than a year later—it is crucial to revisit these findings, using a larger dataset to assess whether the database has improved in consistency, completeness, and overall transparency, or whether the initial concerns remain unresolved. These issues are particularly relevant during a major continental election, in which moderation practices may experience major changes.

Data

Our dataset consists of 1.58B SoRs that we collected from the publicly-available DSA-TDB.² The SoRs cover all self-reported moderation actions that eight very large social media platforms took in the EU between March and October 2024. The chosen time frame covers approximately 14 weeks before and 20 weeks after the election days, allowing a thorough analysis of the possible shift in moderation practices before and after the electoral period. In detail, our data includes 646.1M SoRs from TikTok, 300.8M from Instagram, 260.2M from Facebook, 81.9M from Pinterest, 36.3M from YouTube, 2.3M from Snapchat, 628K from X, and 293K from LinkedIn. As explained in the database’s official documentation, each SoR is composed of multiple fields.³ All those utilized in this study are described in Appendix Table 1.

Analyses and Results

Trends in moderation actions

To address RQ1, we first examine whether platforms adjusted the volume or timeliness of their moderation actions in response to the heightened integrity risks of the 2024 EU elections. A surge in moderation activity or a reduction in moderation delays could indicate increased vigilance, whereas stationarity in these metrics might suggest that moderation remained unchanged despite the elections. To analyze these trends, we constructed daily time series of moderation decisions made by each platform. Additionally, we computed daily time series of SoR moderation delays,

²<https://transparency.dsa.ec.europa.eu/explore-data/download>

³<https://transparency.dsa.ec.europa.eu/page/documentation>

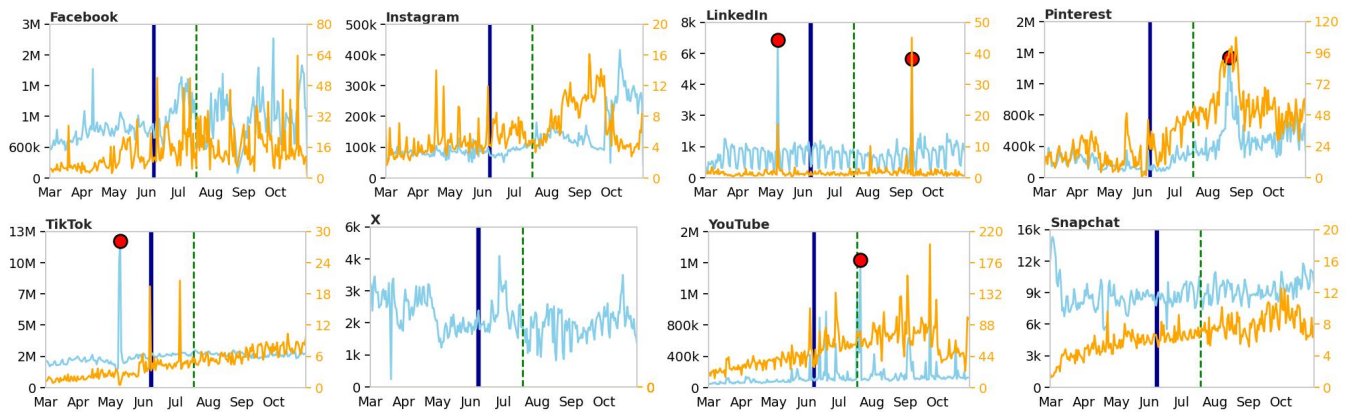


Figure 1: Daily time series of the number of moderation actions (cyan-colored) and their average delay (orange-colored). The blue vertical band indicates the Parliament elections days (6th–9th June), while the green vertical dashed line indicates the Presidential election day (18th July). The red circles highlight the subset of anomalies that we analyzed in detail.

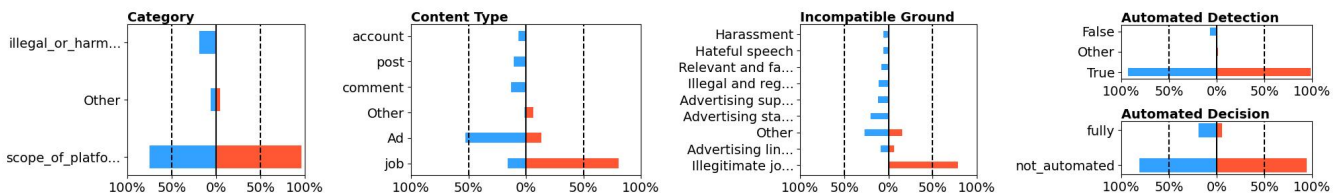


Figure 2: Comparison between the SoRs (right-aligned, red-colored) that caused the moderation *volume* anomaly reported by **LinkedIn** on **May 8, 2024** and the SoRs (left-aligned, blue-colored) by the same platform from the surrounding routine days.

defined as the number of days between the content’s creation of the SoR and its eventual moderation action. For each platform, moderation delays were averaged on a daily basis. Figure 1 presents both the volume (cyan-colored) and delay (orange-colored) time series, allowing us to assess whether notable shifts in moderation practices occurred during the electoral period compared to routine periods.

Figure 1 reveals substantial heterogeneity across platforms in both moderation volume and delay patterns. Some platforms take millions of moderation actions daily, while others only a few thousands. Certain platforms, like X, show stable trends in moderation delay, while others, like LinkedIn, have regular fluctuations in moderation volume. Moreover, some platforms present frequent and sharp variations, indicating dynamic or event-driven moderation policies. The differences persist even when accounting for each platform’s number of active users (Trujillo, Fagni, and Cresci 2025). However, no clear or systematic shift in moderation behavior is observed during the electoral period, indicated by the vertical blue and green lines corresponding to the European Parliament and Presidential elections, respectively. This suggests that, at a broad level, moderation activity remained relatively consistent before, during, and after the elections. Nonetheless, nearly all platforms exhibit distinct spikes in either the volume or delay of moderation at various points in time. To determine if these anomalies are linked to election-related tampering, we conduct a focused analysis on a subset of these peaks. The anomalies selected

for further scrutiny are marked with red dots in Figure 1.

Anomalies in moderation actions

The time series analysis revealed no clear shifts in moderation trends. However, the presence of sharp spikes in moderation volume or delay might indicate that changes have occurred in a more localized manner. For example, these anomalies could indicate mass election-related moderation events, such as coordinated enforcement actions or ban waves, that are not reflected in the broader trends (DeCook 2022). To investigate the nature of the selected anomalies, we compared the characteristics of the SoRs within each spike to those issued by the same platform in the days before and after the spike. This comparative analysis aims to identify possible marked shifts in the attributes of the SoRs that could explain the underlying moderation decisions. To this end, we are particularly interested in the use of specific SoR attributes and values designed to indicate election-related tampering (e.g., the predefined category `negative_effects_on_civic_discourse_or_elections`). Among the available ones, we analyze the type of infringement (`category`), the type of moderated content (`content_type`), the specific reason for incompatibility with platform policies (`incompatible_ground`), and the use of automation in moderation—distinguishing between `automated_detection` and `automated_decision`. For clarity and brevity, in the following analyses we show a

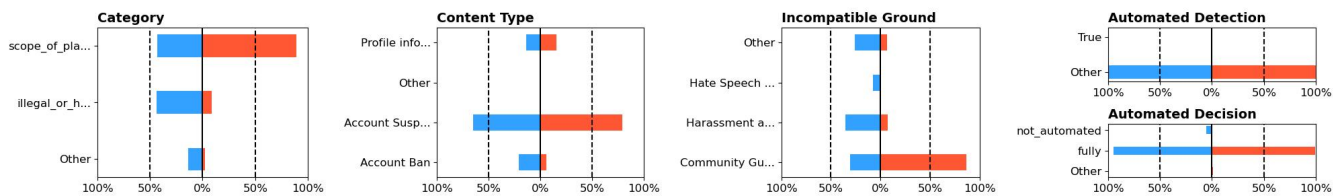


Figure 3: Comparison between the SoRs (right-aligned, red-colored) that caused the moderation *volume* anomaly reported by **TikTok** on **May 10, 2024** and the SoRs (left-aligned, blue-colored) by the same platform from the surrounding routine days.

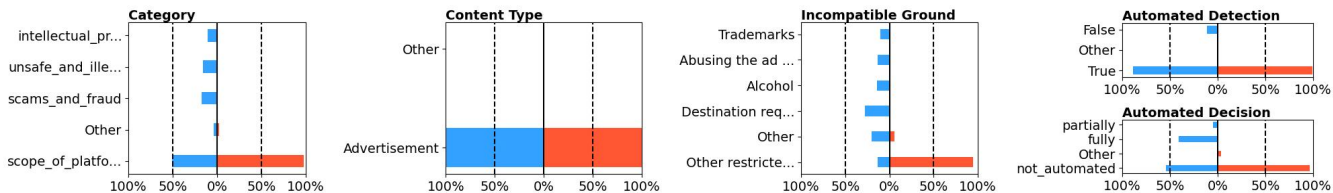


Figure 4: Comparison between the SoRs (right-aligned, red-colored) that caused the moderation *volume* anomaly reported by **YouTube** on **July 21, 2024** and the SoRs (left-aligned, blue-colored) by the same platform from the surrounding routine days.

subset of all attribute values—those mostly used by each platform and those for which we measured noticeable differences. Appendix Section “Attributes values” provides more information on attribute value selection. For each one, we visualize the differences using diverging bar charts, where left-aligned blue bars represent SoRs from routine (non-peak) days, and right-aligned red bars correspond to SoRs associated with the anomalies.

LinkedIn—May 8, 2024. Figure 2 reveals that the SoRs associated with this anomaly differ significantly from routine moderation cases, particularly in the type of moderated content and the grounds for infringement. While LinkedIn mostly moderates advertisements, on May 8, 2024, the platform primarily targeted illegitimate job offers, suggesting a distinct shift in enforcement focus on that day. However, the available data do not provide further details on the rationale behind the moderation of such job offers, nor do they offer clear indications that would allow referring this spike to the electoral context.

TikTok—May 10, 2024. Figure 3 indicates that this anomaly stems from the mass suspension of accounts deemed to be operating outside of TikTok’s intended scope and in violation of its community guidelines. However, the information provided in these SoRs is highly generic, offering little insight into the specific reasons behind this moderation surge or whether it is directly linked to the electoral context.

YouTube—July 21, 2024. Figure 4 shows that this moderation spike primarily targeted advertisements flagged for being outside YouTube’s scope and, specifically, related to “other restricted businesses.” Furthermore, unlike YouTube’s usual moderation processes, all decisions in this case were issued automatically. Interestingly, some type of restricted business could be election-related,⁴ such as businesses related to “Government documents and official ser-

vices.” However, the SoRs submitted by YouTube do not specify which type of restricted businesses they are related to. As such, akin to the previous anomalies, the limited information provided makes it difficult to determine whether the moderation actions were directly related to the election.

Pinterest—August 21, 2024. Figure 5 surfaces minimal differences between the SoRs related to the moderation spike and the regular ones. The few differences show more actions against graphic violence rather than pornographic content. These SoRs suggest this spike wasn’t election-related.

LinkedIn—September 11, 2024. Figure 6 presents the most compelling anomaly in our analysis, revealing a moderation spike driven by a surge in comments flagged for election-related misinformation. A key indicator of its relevance is the explicit use of the `negative_effects_on_civic_discourse_or_elections` field—an unusually specific designation compared to the more generic moderation categories typically employed. Additionally, this anomaly stands out due to its lower reliance on automation, suggesting a more deliberate review process. We also note that this spike appears in the moderation delay time series, indicating that the moderated content was not recent but had been posted weeks earlier. Although the moderation actions occurred in September—after the electoral period—the delay of approximately 45 days traces the original publication of the moderated comments back to mid July. This timing is highly significant, as it falls between the European Parliament and Presidential elections, a period of heightened political discourse and potential misinformation risks. The distinct characteristics of this anomaly, both in terms of timing and the specificity of the moderation labels used, strongly suggest that this spike was indeed election-related. This case highlights the importance of granular labeling practices within the DSA-TDB, as LinkedIn’s use of precise categories rather than broad classifications (e.g.,

⁴<https://support.google.com/adspolicy/answer/6368711?hl=en>

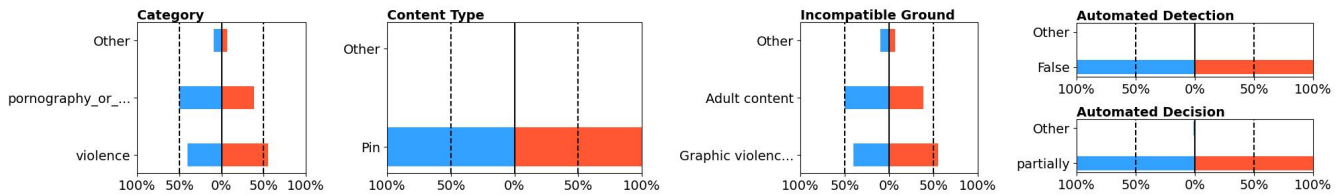


Figure 5: Comparison between the SoRs (right-aligned, red-colored) that caused the moderation *volume* anomaly reported by **P** Pinterest on **August 21, 2024** and the SoRs (left-aligned, blue-colored) by the same platform from the surrounding routine days.

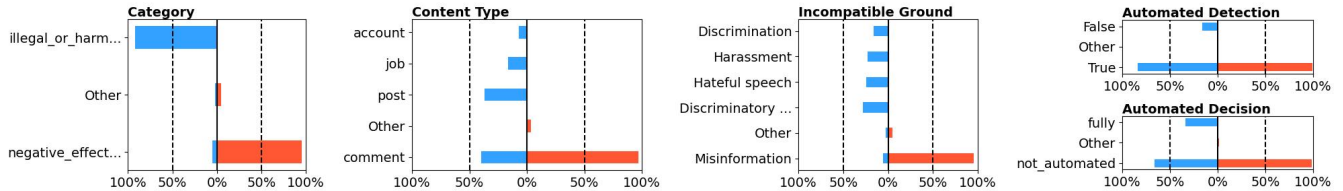


Figure 6: Comparison between the SoRs (right-aligned, red-colored) that caused the moderation *delay* anomaly reported by **l** LinkedIn on **September 11, 2024** and the SoRs (left-aligned, blue-colored) by the same platform from the surrounding routine days.

scope_of_platform_service), enabled us to draw this conclusion with confidence.

Overall, our analysis highlights a significant limitation in assessing the observed moderation anomalies in relation to the European elections. In most cases, we were unable to confidently determine whether these surges in moderation actions were linked to the electoral context or driven by other factors. This uncertainty primarily stems from the lack of detailed information in the SoRs, which manifests in two key ways: (i) the frequent use of generic values in mandatory fields, and (ii) the underutilization of optional fields, which could have provided crucial additional context.

Delays in moderation actions

The analysis of moderation trends revealed no significant changes in moderation volume during the electoral period. However, the orange-colored moderation delay time series shown in Figure 1 reveal a steady increase in average daily delay for all platforms, except for X and LinkedIn. The increase is particularly evident in the post-electoral period, starting in late August and early September 2024. This pattern suggests that while moderation activity remained steady during the elections, platforms may have begun retrospectively moderating election-period content only after the elections had concluded. This delayed response could explain both the absence of noticeable shifts during the electoral period and the rise in moderation delays afterward. To investigate this hypothesis, we conducted a deeper analysis of moderation delays, examining whether content posted during the elections was moderated at a later stage.

Figure 7 presents the joint and marginal distributions of moderation date (x axis) and content publication date (y axis) for each platform. The central heatmaps show the relationship between these dates, on a logarithmic scale. Points along the main diagonal indicate no moderation delay (i.e.,

content moderated on the same day it was published), while points further below the diagonal represent moderation actions targeting older content. In this visualization, no point should ever lay above the main diagonal, as that would correspond to content moderated before being published. Figure 7 reveals highly heterogeneous moderation behaviors across platforms. LinkedIn and X consistently moderate content close to its publication date, with X reporting zero moderation delay in all its submitted SoRs. YouTube, Instagram, and Snapchat primarily focus on recent content but occasionally moderate older posts. In contrast, Facebook and TikTok display a more uniform distribution, with moderation delays more evenly spread across time. Pinterest stands out, showing a moderation pattern largely independent of publication date, as indicated by the vertical lines in its heatmap. Other notable patterns are the dark-colored diagonal lines visible in Facebook’s and Snapchat’s heatmaps. Facebook, in particular, shows a consistent one-month lag for certain moderation actions, possibly reflecting batch reviews or scheduled automated moderation processes. Similar yet milder patterns appear in Instagram and YouTube, hinting at structured moderation processes.

Beyond these observations, Figure 7 allows us to test our initial hypothesis. If platforms had disproportionately moderated content after the electoral period, we would expect to see a clear pattern in the heatmaps—namely, a concentration of moderation actions targeting content published around the Parliament and President elections, but occurring predominantly from early September onward. However, no such pattern emerges in the data. The only notable exception is LinkedIn, where moderation of election-related misinformation comments left a visible mark in the platform’s heatmap. Apart from this case, we find no evidence that moderation actions systematically targeted electoral-period content after the elections had concluded. Thus, with the exception of

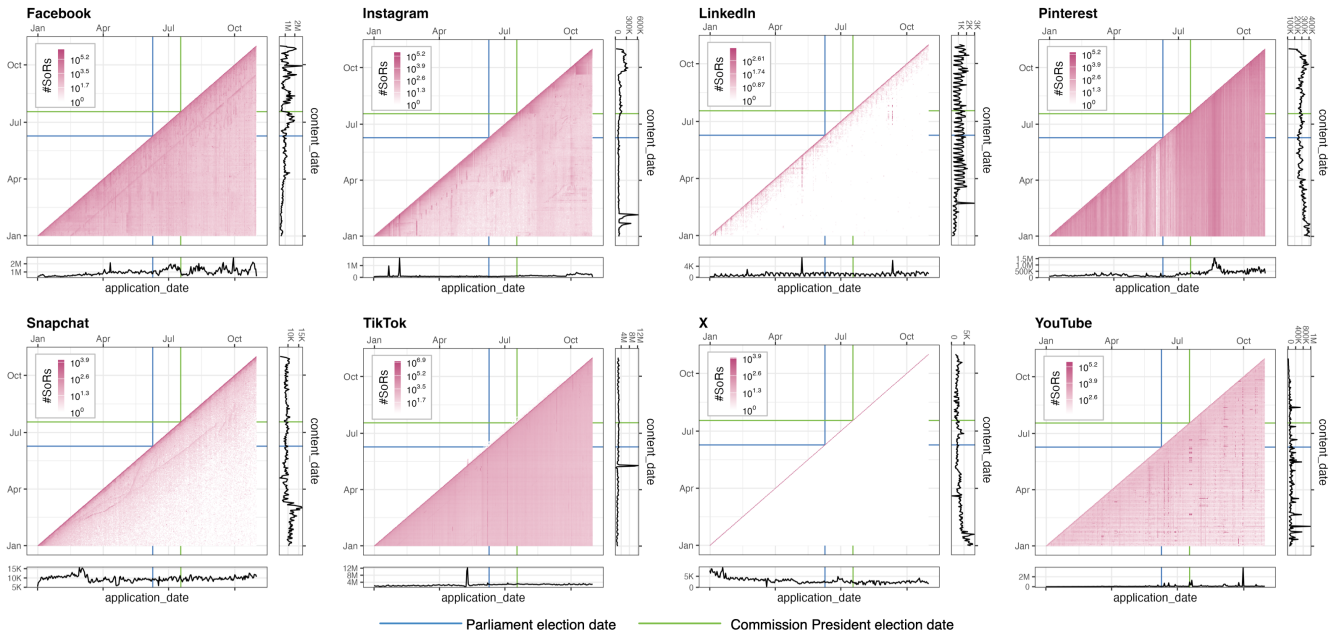


Figure 7: Analysis of moderation delays. For each platform, the heatmaps show the relationship, on a logarithmic scale, between the date when content was moderated (x axis) and the date when the same content was published (y axis). Blue lines indicate the Parliament elections days, while green lines indicate the President election day.

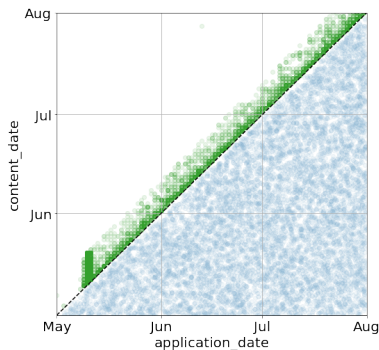


Figure 8: Highlight of 12,010 SoRs from TikTok that erroneously report content being moderated before it was published. The erroneous SoRs are green-colored while the correct ones are blue-colored.

LinkedIn, our analyses of moderation trends, anomalies, and delays do not indicate any significant shifts in moderation practices before, during, or after the electoral period across the analyzed platforms.

Quality of platform-shared data

In the months following the launch of the DSA-TDB, multiple studies uncovered important issues with the quality and consistency of the data submitted by platforms (Trujillo, Fagni, and Cresci 2025; Kaushal et al. 2024; Drolsbach and Pröllochs 2024; Papaevangelou and Votta 2024). Given that the database was designed to promote transparency and accountability, these early shortcomings raised concerns about its reliability. Now—one year after its implementation—

we revisit these findings by analyzing our recent dataset to answer RQ2 and determine whether the situation has improved. If the same inconsistencies persist, this would call into question not only the accuracy of the reported data but also the broader effectiveness of the database as a tool for regulatory oversight.

Erroneous database records. To assess the quality and reliability of the data, we first searched for obvious errors. Our analysis uncovered two issues: (i) duplicate records, and (ii) records indicating that content was supposedly moderated before it was even published. The first issue affected a small minority of records submitted by Facebook and Pinterest, and involves SoRs with identical Universally Unique Identifiers (UUIDs), which should be unique by design. In every case we identified, these duplicate SoRs were completely identical across all fields. The second issue is more concerning: several SoRs by TikTok, Facebook, Snapchat, and LinkedIn reported moderation actions occurred before the corresponding content was published. Figure 8 highlights this pattern in a subset of TikTok data from May to August 2024, where green-colored records erroneously indicate preemptive moderation. Although these errors represent a small fraction of the overall SoRs—for example, TikTok submitted over 12k erroneous SoRs on May 10, 2024, accounting for $\sim 0.1\%$ of its total that day—care is needed when analyzing the data, as filtering decisions could disproportionately retain the flawed records, making them a non-negligible share of the analyzed subset. While the duplication issue likely stems from errors within the DSA-TDB itself, the incorrect moderation timestamps indicate reporting failures on the platforms' side. These errors persist in the database despite being easily detectable, suggesting a lack

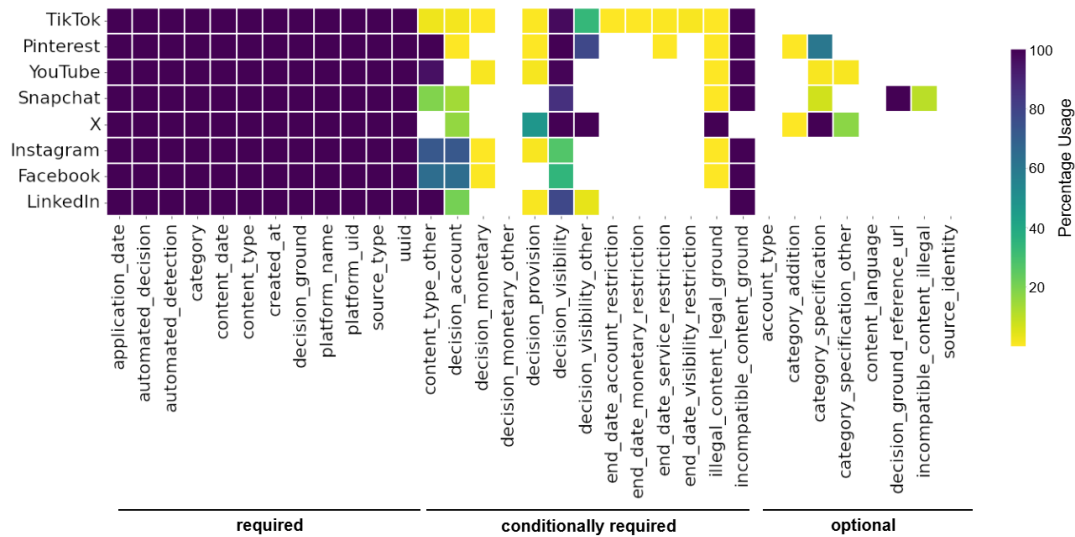


Figure 9: Heatmap showing the frequency with which each database attribute is used by each platform. Attributes are grouped based on whether they are required, conditionally required, or optional. Platforms are displayed in descending order of attribute utilization.

of enforcement at the infrastructure level. Thus, researchers and policymakers must run integrity checks to prevent misleading conclusions (Tessa et al. 2025).

Fields usage and uninformative reporting. The DSA-TDB is structured around required, conditionally required, and optional fields, providing a standardized yet flexible reporting framework. Additionally, certain fields require selecting from a predefined set of values while others allow free-text input. Among the predefined values, some are highly specific while others are broad and generic. Based on this structure, the usefulness of the database hinges on how the platforms populate these fields.

To assess the informativeness of the submitted SoRs, we analyzed the frequency with which each database attribute is used by the different platforms, distinguishing between required, conditionally required, and optional. Figure 9 presents our findings. As expected, all platforms consistently populate the required attributes, ensuring formal compliance with the DSA. However, the use of conditionally required attributes varies widely. While TikTok utilizes most of these attributes, the majority of platforms use almost none. Interestingly, some conditionally required attributes—such as `decision_visibility` and `incompatible_content_ground`—are consistently valued across all platforms, whereas others—such as `decision_monetary_other`—remain entirely unused. Figure 9 also reveals that the optional attributes are overwhelmingly neglected. TikTok, Instagram, Facebook, and LinkedIn never populated any of them, while Pinterest, YouTube, Snapchat, and X used only a handful, in a small minority of SoRs. This issue was already discussed in early assessments of the DSA-TDB (Trujillo, Fagni, and Cresci 2025), and our analysis confirms the lack of progress in addressing it.

One of the required attributes in the DSA-TDB man-

dates platforms to specify the type of infringement leading to a moderation action, using a set of predefined values. While some predefined categories are quite precise (e.g., pornography, harmful speech), others, such as `scope_of_platform_service`, are vague catch-all labels encompassing a wide range of age, geographical, and language restrictions, disallowed goods and services, and nudity.⁵ Early studies noted a frequent reliance on this generic category, raising concerns about the clarity of platforms’ reporting (Trujillo, Fagni, and Cresci 2025; Kaushal et al. 2024). To assess whether this practice has changed, we compared the usage of the `scope_of_platform_service` category in the first 100 days of the database (353M SoRs from September 25, 2023 to January 2, 2024) with the last 100 days of our observed period (573M SoRs from July 24 to October 31, 2024). Here, we refer to the former as the *initial* period and to the latter as the *latest* period. Figure 10 presents the results of this comparison. On average, the use of this category has remained practically unchanged, decreasing only slightly from 42.68% to 41.04%. However, platform-specific usage varied. Facebook, Instagram, LinkedIn, and X have increased their reliance on this generic label, while YouTube and Snapchat have significantly reduced it. TikTok and Pinterest have shown little change. These findings confirm that platforms continue to rely heavily on vague classifications, limiting the informativeness of mandatory fields. This result, combined with the rare use of optional fields, suggests that while platforms meet the formal DSA requirements, the transparency and usefulness of their reporting remain limited.

Unresolved issues in X’s reporting. In addition to identifying widespread issues with the database, previous works

⁵<https://transparency.dsa.ec.europa.eu/page/documentation#16-category-specification-category-category-addition-category-specification>

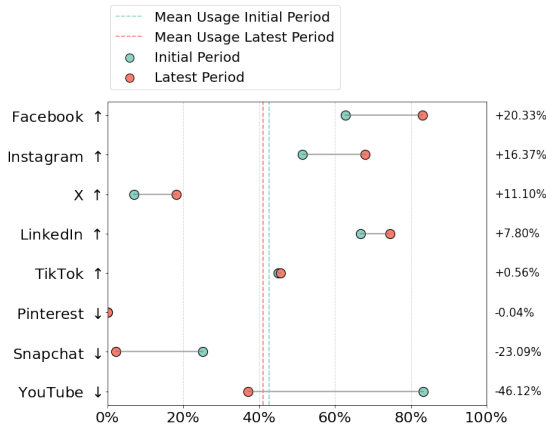


Figure 10: Frequency of platforms’ use of the generic category `scope_of_platform_service` in their SoRs. Teal dots show usage in the initial period and red dots in the latest. Horizontal bars indicate changes between periods. Platforms are ordered by descending difference between the periods. Arrows beside platform names indicate increases (↑) and decreases (↓) in frequency of use. Vertical dashed lines show mean values for each period.

also highlighted some platform-specific limitations. In particular, X stood out as the platform that presented the most inconsistencies in their reporting (Trujillo, Fagni, and Cresci 2025; Kaushal et al. 2024; Drolsbach and Pröllochs 2024). Unlike all other platforms where low moderation delays are consistently linked to a high reliance on automation (Papaevangelou and Votta 2024), X continues to report near-instantaneous moderation actions while claiming to rely exclusively on manual detection and decision-making. Given that deepfakes—X’s almost exclusive target of moderation—are inherently difficult to identify manually at scale, this reporting pattern seems implausible. Further compounding the issue, past research showed that X’s DSA-TDB submissions contradict its own transparency reports, raising concerns about the integrity of its reports (Trujillo, Fagni, and Cresci 2025). Notably, this lack of transparency was one of the factors that led the European Commission to open formal proceedings against X on December 2023 (European Commission 2023). To assess whether X has addressed these inconsistencies, we compared its reporting from the initial and latest period. As Figures 1 and 7 illustrate, X consistently reported zero moderation delay throughout the entire period. Additionally, Figure 11a shows that 99% of X’s recent SoRs continue to indicate purely manual moderation, while Figure 11b confirms its sustained focus on synthetic media. The minimal variations between the two periods are statistically non-significant ($p = 0.99$, χ^2 test), indicating no meaningful changes in X’s reporting practices. These findings confirm initial concerns about X’s data reliability and show no improvement over time.

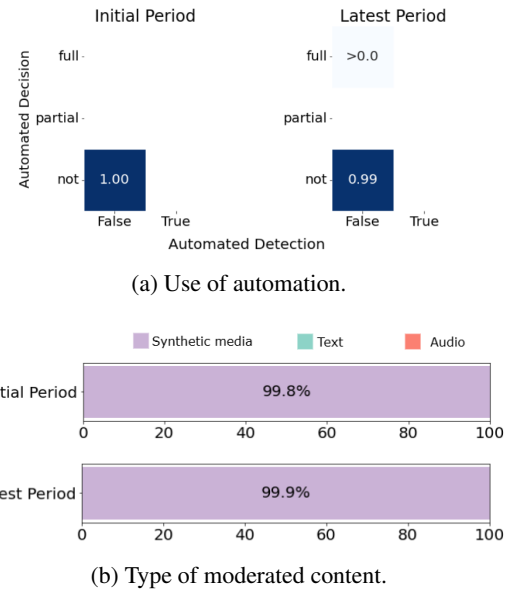


Figure 11: Comparison between the use of automation (a) and the types of moderated content (b) reported by X during the initial and latest periods.

Discussion and Conclusions

We analyzed 1.58B moderation actions on the the Digital Services Act Transparency Database (DSA-TDB)—a tool designed to enhance transparency and observability in online moderation—to shed light on how major social media platforms handled content moderation during the 2024 European Parliament elections—a large-scale, multi-country political event. In RQ1 we sought to determine whether the DSA-TDB revealed changes in content moderation practices before, during, or after the electoral period. Apart from LinkedIn’s delayed enforcement against election-related misinformation, we found no evidence of meaningful shifts in moderation behavior across the eight analyzed social media platforms. This finding bears important implications. One way to explain the result is that platforms simply did not adjust their moderation strategies in response to the heightened integrity risks that are expected during political events of such magnitude (Pierri et al. 2023; Majó-Vázquez et al. 2021). If true, this finding suggests that platforms either deemed their existing moderation frameworks sufficient or chose not to implement election-specific interventions. Alternatively, platforms may have indeed modified their enforcement practices, but the data available on the DSA-TDB was insufficient to reveal such changes. This could stem from two intertwined limitations: (i) self-reporting deficiencies, and (ii) structural shortcomings in the database. Since the DSA-TDB relies on voluntary submission, platforms may have omitted some details about their enforcement actions, either unintentionally due to internal reporting gaps, or deliberately to maintain a degree of opacity on their moderation practices. At the same time, even if platforms had fully reported their moderation actions, the database’s design may have inherently

prevented the detection of moderation shifts (Trujillo, Fagni, and Cresci 2025). For example, the predefined categories and mandatory fields might be too generic or too limited to capture meaningful variations in enforcement, making it difficult to track how platforms respond to evolving systemic risks like election-related disinformation.

In RQ2, we assessed the state of the DSA-TDB approximately one year after its launch, revisiting the initial findings that highlighted important limitations in the database and evaluating whether these issues had been mitigated or resolved. Our analysis confirms that the same shortcomings—such as incomplete reporting, vague categorization, and unreliable data, particularly from X—persist without significant improvement. This stagnation raises concerns not only about the database’s ability to serve as a robust transparency tool but also about the broader effectiveness of regulatory efforts aimed at increasing accountability in platform governance (Kausche and Weiss 2024). Transparency mechanisms like the DSA-TDB are only as valuable as the quality of the data they provide. If platforms systematically underuse informative fields, rely on too generic classifications, or submit records that defy plausibility, then the promise of meaningful oversight is undermined. Beyond regulatory scrutiny, these findings also speak to the role of transparency in public trust. Research has shown that users are more likely to accept and support moderation decisions when platforms provide clear, well-documented rationales (Cai et al. 2024; Jhaver, Bruckman, and Gilbert 2019). Yet, the persistent opacity in platform disclosures might suggest a reluctance to fully embrace transparency, potentially reinforcing skepticism toward content moderation practices. If platforms fail to provide detailed, high-quality reports even under legal mandate, this calls into question the limits of transparency-by-design approaches and the need for stronger enforcement mechanisms to ensure compliance.

In conclusion, the implications of our study are multi-fold and relevant. If platforms did not adjust their moderation actions during a high-stakes election, this raises concerns about their responsiveness to systemic risks. Notably, the European Commission has already opened formal proceedings against Facebook, Instagram, and X over deficiencies in mitigating threats to civic discourse and electoral integrity, with specific reference to the 2024 European Parliament elections (European Commission 2024, 2023). This regulatory scrutiny underscores the urgency of ensuring effective and transparent platform moderation practices, particularly in moments of heightened political sensitivity. Conversely, if the structure of the DSA-TDB prevented the surfacing of such shifts, its effectiveness as a transparency tool would be called into question. In either case, our findings suggest that the database, at least in its current form, may not yet fulfill its potential as a mechanism for scrutinizing platform behavior during politically sensitive periods.

Limitations

Our study relies on the quality of the DSA-TDB data, which is subject to platforms’ voluntary self-reporting. The accuracy and completeness of the data depend on the platforms’ willingness and ability to provide detailed information, po-

tentially leading to inconsistencies or gaps in the reported actions. Additionally, our analysis was limited by the set of platforms considered in the study. While we focused on the eight major social media platforms in the EU, our findings may not fully represent the content moderation practices of other platforms. The lack of contextual information in the DSA-TDB presents further challenges. The database provides only metadata on moderation actions, without revealing the actual pieces of content that were moderated. As a result, it is impossible to assess the exact nature or context of the moderated content. This limitation—while necessary to protect user privacy—restricts the depth of our analyses. Finally, it remains unclear whether the absence of visible changes could be due to limited discourse around the European elections, resulting in little content to moderate.

Future works

One of the key limitations of our study was the inability to directly analyze the content that was moderated, due to the lack of content identifiers. To this end, ongoing initiatives by the European Commission aimed at designing working procedures for access to platform data under Article 40 of the DSA, could relieve the issue.⁶ When these become available, they could provide an opportunity to combine the self-reported records from the DSA-TDB with the corresponding platform data, allowing for a much richer and complete analysis of content moderation practices. Looking ahead, it would also be valuable to assess potential changes in how platforms report their moderation actions following the ongoing formal proceedings against TikTok, Facebook, Instagram, and X, which may prompt improvements in reporting practices. Furthermore, ensuring the ongoing quality of data in the DSA-TDB will be crucial for future transparency efforts, and our analysis could be revisited to evaluate the evolution of platform moderation during other major events, whether political or otherwise.

Acknowledgments

This work is partly supported by the ERC project DEDUCE under grant #101113826 and by the European Union – Next Generation EU, Mission 4 Component 1, for project PIANO (CUP B53D23013290006).

References

- Aspromonte, M.; Ferraris, A.; Galli, F.; and Contissa, G. 2024. LLMs to the Rescue: Explaining DSA Statements of Reason with Platform’s Terms of Services. In *NNLP*.
- Bär, D.; Pierri, F.; De Francisci Morales, G.; and Feuerriegel, S. 2024. Systematic discrepancies in the delivery of political ads on Facebook and Instagram. *PNAS Nexus*, 3(7).
- Bene, M.; Ceron, A.; Fenoll, V.; Haßler, J.; Kruschinski, S.; Larsson, A. O.; Magin, M.; Schlosser, K.; and Wurst, A.-K. 2022. Keep them engaged! Investigating the effects of self-centered social media communication style on user engagement in 12 European countries. *Political Communication*.

⁶https://www.eu-digital-services-act.com/Digital_Services_Act_Article_40.html

- Cai, J.; Patel, A.; Naderi, A.; and Wohn, D. Y. 2024. Content moderation justice and fairness on social media: Comparisons across different contexts and platforms. In *ACM CHI*.
- Chen, L.; Chen, J.; and Xia, C. 2022. Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64: 103060.
- Cima, L.; Miaschi, A.; Trujillo, A.; Avvenuti, M.; Dell’Orletta, F.; and Cresci, S. 2025. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *ACM WebConf*.
- Cinelli, M.; Cresci, S.; Galeazzi, A.; Quattrociocchi, W.; and Tesconi, M. 2020. The limited reach of fake news on Twitter during 2019 European elections. *PLoS One*, 15(6).
- Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2014. A criticism to society (as seen by Twitter analytics). In *IEEE ICDCS Workshops*.
- DeCook, J. R. 2022. r/WatchRedditDie and the politics of Reddit’s bans and quarantines. *Internet Histories*, 6(1-2).
- Dergacheva, D.; Kuznetsova, V.; Scharlach, R.; and Katzenbach, C. 2023. One day in content moderation: Analyzing 24h of social media platforms’ content decisions through the DSA Transparency Database. Technical report, Lab Platform Governance, Media, and Technology (PGMT). Centre for Media, Communication and Information Research (ZeMKI), University of Bremen.
- Diakopoulos, N.; and Johnson, D. 2021. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7).
- Drolsbach, C. P.; and Pröllochs, N. 2024. Content moderation on social media in the EU: Insights from the DSA Transparency Database. In *ACM WebConf Companion*.
- European Commission. 2023. Commission opens formal proceedings against X under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_6709. Accessed: 25 March 2025.
- European Commission. 2024. Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act. <https://ec.europa.eu/commission/presscorner/detail/en/ip.24.2373>. Accessed: 25 March 2025.
- European Parliament and Council. 2022. Regulation on a Single Market For Digital Services (Digital Services Act) and amending Directive. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065>.
- Fatema, S.; Yanbin, L.; and Fugui, D. 2022. Social media influence on politicians’ and citizens’ relationship through the moderating effect of political slogans. *Frontiers in Communication*, 7.
- Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Haq, E.-U.; Zhu, Y.; Hui, P.; and Tyson, G. 2024. History in making: Political campaigns in the era of artificial intelligence-generated content. In *ACM WebConf Companion*.
- Hua, Y.; Naaman, M.; and Ristenpart, T. 2020. Characterizing Twitter users who engage in adversarial interactions against political candidates. In *ACM CHI*.
- Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. In *ACM CSCW*.
- Kausche, K.; and Weiss, M. 2024. Platform power and regulatory capture in digital governance. *Business and Politics*.
- Kaushal, R.; Van De Kerkhof, J.; Goanta, C.; Spanakis, G.; and Iamnitich, A. 2024. Automated Transparency: A legal and empirical analysis of the Digital Services Act Transparency Database. In *ACM FAccT*.
- Kim, D. H.; and Ellison, N. B. 2022. From observation on social media to offline political participation: The social media affordances approach. *New Media & Society*, 24(12).
- Majó-Vázquez, S.; Congosto, M.; Nicholls, T.; and Nielsen, R. K. 2021. The role of suspended accounts in political discussion on social media: Analysis of the 2017 French, UK and German elections. *Social Media + Society*, 7(3).
- Matatov, H.; Naaman, M.; and Amir, O. 2022. Stop the [Image] steal: The role and dynamics of visual content in the 2020 US election misinformation campaign. In *ACM CSCW*.
- Mazza, M.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2022. Investigating the difference between trolls, social bots, and humans on Twitter. *Computer Communications*, 196.
- Papaevangelou, C.; and Votta, F. 2024. Content moderation and platform observability in the Digital Services Act.
- Papakyriakopoulos, O.; Engelmann, S.; and Winecoff, A. 2023. Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. In *ACM CHI*.
- Pierri, F.; Luceri, L.; Chen, E.; and Ferrara, E. 2023. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Science*, 12(1).
- Rho, E. H. R.; and Mazmanian, M. 2020. Political hashtags & the lost art of democratic discourse. In *ACM CHI*.
- Shahi, G. K.; Basyurt, A. S.; Stieglitz, S.; and Neuberger, C. 2024. Agenda formation and prediction of voting tendencies for European parliament election using textual, social and network features. *Information Systems Frontiers*.
- Shahi, G. K.; and Mejova, Y. 2025. Too little, too late: Moderation of misinformation around the Russo-Ukrainian conflict. In *ACM WebSci*.
- Tardelli, S.; Avvenuti, M.; Tesconi, M.; and Cresci, S. 2020. Characterizing social bots spreading financial disinformation. In *HCI*.
- Tessa, B.; Amram, D.; Monreale, A.; and Cresci, S. 2025. Improving regulatory oversight in online content moderation. In *HAI Workshops*.
- Trujillo, A.; Fagni, T.; and Cresci, S. 2025. The DSA Transparency Database: Auditing self-reported moderation actions by social media. In *ACM CSCW*.

Ethics Checklist

1. General items

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **No, see [Limitations on why this was not feasible](#)**.
- (e) Did you describe the limitations of your work? **Yes**
- (f) Did you discuss any potential negative societal impacts of your work? **NA**
- (g) Did you discuss any potential misuse of your work? **NA**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**

2. Hypotheses testing.

- (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
- (b) Have you provided justifications for all theoretical results? **NA**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
- (e) Did you address potential biases or limitations in your theoretical framework? **NA**
- (f) Have you related your theoretical results to the existing literature in social science? **NA**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**

3. Theoretical proofs.

- (a) Did you state the full set of assumptions of all theoretical results? **NA**
- (b) Did you include complete proofs of all theoretical results? **NA**

4. Machine learning experiments.

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**

- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
- (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **NA**

5. Use of existing assets (e.g., code, data, models) and release of new assets.

- (a) If your work uses existing assets, did you cite the creators? **Yes**
- (b) Did you mention the license of the assets? **NA**
- (c) Did you include any new assets in the supplemental material or as a URL? **NA**
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **Yes, the data we used is freely downloadable. See [Data section](#)**.
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **NA**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **NA**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **NA**

6. Crowdsourcing and research with human subjects.

- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
- (d) Did you discuss how data is stored, shared, and de-identified? **NA**

Appendix

Attributes description

Although the DSA Transparency Database provides a wide range of attributes and information, we focus our analysis on a subset of the most relevant attributes for our study. The complete list of analyzed attributes, along with references to the corresponding sections of the official documentation, and brief descriptions, is provided in Table 1.

Attributes values

Platforms can assign a predefined set of values to the attributes `category`, `automated_decision`, and `automated_detection`. Therefore, in our analysis of moderation anomalies, we report all possible values for the automation-related attributes, as there are only two for `automated_decision` and three for `automated_detection`. Conversely, for the `category` attribute—which allows up to 14 values—we only include those used at least once, for the sake of brevity and clarity. The `content_type` attribute also requires the use of a predefined set of values, including the value `other`. When a platform only used `other` as the content type, we examined the `content_type_other` attribute, which is a free-text attribute allowing platforms to specify content that did not fit within the predefined set of values. The specifications used in the `content_type_other` attribute differ from platform to platform. Consequently, we merged `content_type` and `content_type_other` into a single category and reported only the values that were used at least once. The same reasoning applies to `incompatible_content_ground`, whose content is fully up to the platforms and not predefined.

field	reference	description
<code>application_date</code>	§4.1 Application Date	Indicates when a content moderation decision was applied
<code>automated_decision</code>	§10. Automated Decision	Indicates whether the decision to moderate a content was automatic or not
<code>automated_detection</code>	§9. Automated Detection	Indicates whether moderated content was detected automatically or not
<code>category</code>	§16. CATEGORY & SPECIFICATION	Indicates the type of illegality or incompatibility with the platform’s terms of services that led to a content being moderated
<code>content_date</code>	§2.3. Date on which the content was created on the online platform	Indicates when moderated content was created
<code>content_type</code>	§2.1. Type of content affected	Type of the moderated content (e.g. audio, video, image, etc.)
<code>content_type_other</code>	§2.2. Specification of Content Type “Other”	Specification required when content type is “other”
<code>decision_ground</code>	§11. Decision Grounds	Indicates whether the moderated content was deemed allegedly illegal or incompatible with the platform’s terms of service
<code>incompatible_content_illegal</code>	§12.2. Explanation of the applicability of the legal ground	Explains why a specific content has been deemed illegal according to Article 17(3)(d) of DSA
<code>incompatible_content_ground</code>	§13.1. Incompatible Content Grounds	Explains why a specific content has been deemed incompatible with the platform’s terms of service

Table 1: Complete list of the DSA Transparency Database (DSA-TDB) fields analyzed in this study. For each field, we report its name, reference to the official documentation, and brief description.