Content moderation on social media platforms
Prof. Dr. Nils Weidmann, M.A. Mia Nahrgang
M.A. Seminar

Winter semester 2024/2025
24. March 2025

Examining Content Moderation Dynamics Around Elections Using the DSA Transparency Database

David Breukel[1], Iordanis Pantzartzis[2], Hedvika Ruszová[3], Jenny H. Skrogstad[4]

[1]Introduction & conclusion
[2]Results
[3]Literature, theory and hypotheses
[4]Research design

Abstract

Social media shapes our society, and platforms can broadly remove or otherwise moderate online content. Yet, social media platforms have been criticized for their susceptibility to misinformation and opaque content moderation practices, which may even pose risks to democratic processes – particularly during elections. Leveraging the Digital Services Act (DSA) Transparyency Database, this paper examines content moderation dynamics around the 2024 European Parliament election. We study the relative frequency and automation of moderation targeting content deemed harmful to civic discourse or elections. Contrary to expectations, we observe a decrease in the share of such moderation as the election approaches. We also find that the automation gap between content that is harmful to civic discourse and other content narrows before the election. However, results from placebo tests suggest that these patterns may not be linked to the election. These findings provide new insights into the implementation of the DSA and platform behavior around elections.

Keywords: Content moderation; Digital Services Act; European Parliament election; misinformation; social media.

Introduction[1]

[1] *David Breukel*

Social media has become a major source of information for many Europeans, with 37% relying on it as one of their main news sources (European Parliament, 2024a). At the same time, there are growing concerns about misinformation and behavior that is harmful to civic discourse thriving on social media. Notable examples include misleading Brexit reports and Russian interference during European Parliament elections (Greene et al., 2021; Herszenhorn & Volpicelli, 2024; Kottasová, 2024). Social media platforms are not powerless in light of this: They can remove, fact-check, or otherwise moderate online content at a large scale. This content moderation is an important but opaque tool.

Attention is the scarcest resource in politics. The ability of both citizens and politicians to make their voices heard depends on their capacity to capture attention in public discourse. Conversely, failing to do so leads to exclusion from the democratic debate. With growing political importance of social media platforms, this struggle for attention takes center stage with unprecedented clarity. Thus, the stakes of moderation decisions are high – and both the political left and the political right have criticized platforms' decisions (González-Bailón et al., 2023; Soave, 2022; A. Thompson, 2020). A central source of this criticism is the lack of transparency in moderation decisions.

Since platforms' moderation policies are embedded within millions of lines of code and involve hundreds of thousands of decisions made by human moderators, some lack of transparency may not come as a surprise. Yet, social media platforms have also been criticized for intentionally blurring their moderation criteria and decisions (Kang & Satariano, 2024). In 2024, the European Union introduced the Digital Services Act (DSA) to enhance transparency, accountability, and safety in the digital space. Under the DSA, large online platforms are required to disclose their decisions on removing or restricting content and to provide detailed statements of reasons for these actions within the DSA Transparency Database. The introduction of the DSA is particularly relevant in the context of elections, where disinformation and opaque moderation practices can undermine democratic processes. The president of the European Parliament, Roberta Metsola, described the European Parliament election as a crucial test of the DSA's ability to combat destructive narratives, propaganda, and disinformation (European Parliament, 2024b).

In this paper, we ask how elections influence content moderation dynamics on social media platforms, particularly regarding content that has negative effects on civic discourse or elections. This category in the DSA Transparency Database refers to content spreading dis- and misinformation, as well as foreign information manipulation and interference (European Commission, 2024b). Focusing on six major social media platforms, we investigate the relative frequency of moderation decisions targeting such content, as well as the extent to which they are automated vis-à-vis decisions on other content types in the lead-up to and aftermath of the European parliamentary election in 2024. In doing so, we obtain comprehensive and detailed evidence on moderation dynamics for content that may be harmful to the election process.

The paper thus serves as an important step in evaluating the effectiveness of the DSA, as well as platforms' commitment to comply with its objectives. Skepticism about this commitment has resulted in numerous investigations into social media companies (e.g., X and Meta) for DSA violations related to a lack of disinformation moderation (Laaff, 2024). Moreover, our article speaks to numerous gaps in the emerging literature on content moderation. It

is, to the best of our knowledge, the first attempt to provide a systematic, cross-platform analysis of moderation dynamics around elections. This contribution is especially important because social media influence a wide range of political outcomes (Zhuravskaya et al., 2020). Previous research on content moderation generally focuses on single platforms and is limited to moderation dynamics within the United States (US).

Contrary to our expectations, we find that the relative frequency of moderation related to content deemed detrimental to civic discourse or elections decreases as the election approaches. Furthermore, the results do not suggest a drop, but instead an increase in the relative frequency immediately after the election. As time goes on, the share of content that is harmful to civic discourse or elections tends to decrease again. This is surprising in light of previous research suggesting that misinformation becomes more prevalent on social media as elections approach (Ferrara, 2017; Grinberg et al., 2019a). Additionally, our results suggest that the difference in the share of automated moderation of content deemed harmful to civic discourse and the share of automated moderation of other content decreases as the election nears. However, placebo tests raise concerns about the robustness of these findings. When replicating the analysis using alternative content categories that should be unrelated to elections, such as pornography, we observe similar effects. This casts doubt on whether the initial findings are driven by election-related dynamics.

Literature, theory and hypotheses[3]

[3] *Hedvika Ruszová*

The rise of social media over recent years has been accompanied by the spread of disinformation and its potential influence on voters' behavior. Past studies have shown that false information on social media can sway election results, reinforce public polarization, and undermine democratic institutions. (Bovet & Makse, 2019; Grinberg et al., 2019b)

Studies on past elections focus primarily on the US context. Firstly, Grinberg et al. (2019) examined the circulation of fake news on Twitter during the 2016 US presidential election. They found that false information predominantly targeted a disproportionately small, highly engaged group of users, coming prevalently from conservative and far-right circles, as most users were informed via factual news. Bovet and Makse (2019) additionally found that approximately a quarter of the news shared on Twitter during this election was extremely biased or false. (Bovet & Makse, 2019) This study used a comprehensive observational dataset from Twitter's 171 million users.

Zhuravskaya et al. provide a comprehensive review of the current debate on the political effects of social media and the Internet in the field of political economy. Their overview suggests that fake news spreads faster and wider than factual information on social media. Recent studies research this social media phenomenon, particularly around political contexts such as elections. (Zhuravskaya et al., 2020)

The spread of disinformation is intertwined with polarization. Munger et al. (2022) reviewed existing findings on the relationship between social media and political polarization. They stated that even though social media does not necessarily deepen polarization, it amplifies the spread of false information. (Huszár et al., 2022; Munger et al., 2022) The question remains what effective reactions and measures can be implemented to moderate the spread of fake news on social media? (Pierri et al., 2023) For example, Gondález-Bailón et al. conducted a study that saw a sizable decrease in misinformation diffusion in times of high-intensity content moderation around the 2020 US election highlighting the importance of moderation measures. (González-Bailón et al., 2024)

Similiar concerns over the moderation of disinformation have sparked on the European continent. Within the European context, the newly introduced DSA Transparency Database offers a promising tool, as it aims to consolidate all of the content moderation actions made by all large social media platforms. The European Union (EU) has considered these concerns and started implementing proper legislation, which resulted in enforcing the Digital Services Act (DSA) in 2024. This legislative act provides new avenues for empirical research on how disinformation and misinformation are moderated (Trujillo et al., 2023) First studies, however, show that the database lacks many key details on decision-making when moderating the content and inconsistency in enforcement (not only) across regions. (Trujillo et al., 2023) Researchers then have to take into consideration these limitations of the DSA database when working with it.

Current research mostly examines the effects of false claims and information on US elections, particularly on the 2016 and 2020 elections. A systematic analysis of the 2024 EU election using the DSA Transparency Database are still needed. Previous studies mainly focus on only one main platform, such as Twitter or Facebook, sometimes a few more platforms. However, they lack analyses of datasets covering multiple platforms, as the DSA Transparency Dataset now can provide. Nevertheless, the DSA is an important step for empirical research on moderation patterns that can help on uncover gaps in current pracitces done by the platforms and contribute to better-countering disinformation not only in the EU.

Theory [HR]

Gillespie (2018) states that content moderation is an inherent responsibility of social media to do and exposes the myth of social media neutrality. Social media must balance the external requirements of different actors such as governments, advocacy groups, users, or regulatory bodies. These requests can become even more difficult to implement around elections when it is harder for social media platforms to balance the external pressure. The demand for content moderation further increased after the revelations of disinformation operations when governments recognized the need to defend their countries against foreign interference. (Gillespie, 2018)

Schaub and Morisi (2019) analyzed the data that highlighted higher support for the populist Five Star Movement in Italy in 2013 and for Alternative in Germany in 2017 in regions with a broadband Internet connectivity. Their results are further supported by Guriev et al. (2020) who see an increase in support for both right-wing and left-wing populist opposition parties in places with larger Internet coverage. (Guriev et al., 2021; Schaub & Morisi, 2020) Their works propose a theory that the Internet connection fosters the rise of populist parties and encourages them to amplify their presence on social media before elections. (hypothesis 1)

We have witnessed a rise of populist parties winning the European election in 2024 when 60 populist parties in comparison to 40 in 2019 gained representation in the European Parliament. (Ivaldi & Zankina, 2024) Waisbord has found that there is a mutually beneficial relationship between populism and the post-truth communication they implement in an era when facts are less effective than appeals to emotions, beliefs, and identities. This atmosphere leads to the amplification of fake narratives on social media and the findings indicate an assumption that elections may be particularly vulnerable to an increase of fake news on social media. (Waisbord, 2018) The concept of algorithmic amplification refers to how algorithms personalize social media timelines and can amplify certain messages while reducing the visibility of others. (Huszár et al., 2022)

Furthermore, the algorithmic amplification (Huszár et al., 2022; Riemer & Peter, 2021) raises an important

question of to what extent social media also depend on algorithmic curation of posts, which prioritizes engagement on posts over their accuracy (hypothesis 2). This goes hand in hand with political competition before an election when political parties try to boost their content onto their audiences (hypothesis 1). As Huszar et al. found in six out of seven countries they studied, the mainstream political right enjoyed significantly higher algorithmic amplification than the mainstream political left. (Huszár et al., 2022)

The 2016 presidential election in the US showcased that the spread of misinformation on the biggest platforms has often been deliberate. (Gillespie, 2018) The political actors (parties and their candidates) may also feel incentivized to campaign more on social media, as it serves as an effective tool to fundraise more money from voters. Hong in 2013 revealed that social media (Twitter in this case) increased significantly donations to politicians and it has appeared especially true in the case of more extreme political representatives. (Hong, 2013)

Additionally, social media algorithms help share and display more interactive and viral content, which can too create a demand for increased algorithmic moderation. Furthermore, as social media content has become enormous, pure human content moderation is now not well manageable and as a response may call for an increase in automated moderation.

Hypotheses [HR]

Political economy literature proposes hypotheses on the political influence of social media following two main features distinguishing new social media from traditional offline ones. These are low barriers to entry and posting and reliance on user-generated content. The spread of any information is easier due low barriers to entry and the absence of traditional gatekeeping, increasing the volume of false or harmful content. As users share content with unprecedented speed, social media can be easily flooded by misinformation and fake news. Automated accounts and content further enable such manipulation of online space strengthened by algorithmic amplification and the creation of echo chambers. (Zhuravskaya et al., 2020)

Nonetheless, as the theory suggests, elections are especially prone to the spread of fake news and disinformation that primarily serve radical or populist parties. With the rise of populist parties during the European election in 2024, there may be larger amounts of false and harmful content. This increase then could also boost automated moderation as large amounts of harmful content are harder to manage by non-automated moderation. Due to the algorithmic amplification of content and spread of disinformation in the run-up to the election, platforms may increase content moderation before an election and relax its enforcement afterward. Following this assumption, social media will likely rely more on automated moderation the closer the elections are.

Taken together, the latest research suggests that election periods are accompanied by higher risks of disinformation campaigns, algorithmic amplification, and pressures for platforms to increase moderation. Therefore, this research paper examines how elections influence moderation patterns on social media by testing the following hypotheses:

H1: The closer to the election, the share of moderation of content categorized as ”Negative Effects on Civic Discourse or Elections” (NECDE) will increase relative to other content.

H2: The closer to the election, the fully or partially automated moderation of content categorized as ”Negative Effects on Civic Discourse or Elections” (NECDE) increases relative to other content.

Research design[4]

[4] *Jenny H. Skrogstad*

To address our research question, we analyze data from the Digital Services Act (DSA) Transparency Database. According to Article 17 of the DSA, providers of online platforms (online marketplaces, app stores, and social media platforms) are mandated to inform users of content moderation practices and provide justification for their decisions (European Commission, 2024a). From August 2023, DSA rules applied to 'very large online platforms' with over 45 million EU users, expanding to all platforms in February 2024 (Commission, 2025a). Statements of reasons (SoR) must be uploaded to the public database in accordance with DSA Article 24. These statements should include details regarding the nature of the restriction, the rationale behind the decision, and the relevant facts and circumstances that informed the content moderation action. (Commission, 2022). These decisions are categorized according to a predefined system, with mandatory and optional classification specifications (see figure 13 in the Appendix for a overview of the 13 manda categories) (European Commission, 2025). As of March 2025, 150 active platforms adhere to DSA rules, with the most common reported violations of the "Scope of platform service", "Unsafe and/or illegal products", and "Illegal or harmful speech". 49 % of the decisions were fully automated. (Commission, 2025b).

The database is publicly accessible and machine-readable, allowing us to extract the necessary data we need to investigate our research question. We downloaded all SoRs in our period of observation for the platforms we want analyze via the databse web API. The data from the Web API come in the form of several dumps per day, so we first compiled the data into daily files[1]. We further compiled the daily files into one large database and then used datastreaming to wrangle the dataset, which allowed data manipulation of large datasets with finite computer memory.

Our study focuses on the European Parliament election held between June 6 and 9, 2024. For the purpose of this paper we consider June 9 election day. To assess any potential impact on content moderation, we analyze data from 50 days before, during, and 50 days after the election day. As related to our research question, we identify 10 relevant social media platforms in the database: Facebook, Instagram, LinkedIn, Pinterest, Reddit, Snapchat, Threads, TikTok, X, and YouTube. We are mainly interested in statements of reasons for the category "Negative effect on civic discorse or elections" (NECDE). Other relevant categories for our research question could be "Illegal or harmful speech" and "Risk for public security". However, since the database only allows for the selection of one category, NECDE is the most fitting in our case. As related to our second hypothesis we are also interested in information about whether the moderation decision was automated or not. The database categorizes these processes into three distinct types: fully automated decisions, partially automated decisions, and non-automated decisions. A fully automated decision indicates that the entire decision-making process was conducted without any human intervention. In contrast, a non-automated decision signifies that the decision-making process was executed without the assistance of automated tools. The partially automated decision category encompasses processes where both automated methods and human oversight were utilized (European Commission, 2025).

Figure 1 displays the total number of SoRs and the number of SoR of NECDE on a logarithmic scale for each social media platform. The logarithmic scale is used for better visualization, as the number of SoR can reach several millions for one day. The platform Threads is missing from the graphics as it had no SoRs for this timeline.

[1] See the supplemental material for the download script

The social media platform Reddit only has information for the last period of the timeline we are investigating. X and YouTube do not have any SoRs of NECDE, which is the main SoR category we are investigating. It was therefore decided to drop these four platforms from our analysis. The fact that several platforms do not report any violations for the category NECDE can mean that the platforms do not report their SoRs correclty or thourghly, leading to that it can be hard to trust that the data we are using is completely correct. Figure 1 show that SoRs in the NECDE category is quite low for all platforms, especially for the platforms Facebook and Instagram.

Figure 2 shows the overview of the total number of SoR and the number of automated decisions for the 6 included social media platforms. Pinterest is one of the platforms with the most automated decisions, while Snapchat has more human moderated decisions compared to automated decisions. Generally, the data shows a positive correlation between the total number of SoRs and the number of automated SoRs across most platforms—when the total number of SoRs increases, the number of automated SoRs also rises. However, TikTok stands out in the pre-election period, where the number of automated decisions do not follow the same trend as the total number of SoR. Both figure 1 and figure 2 show that TikTok stands out in terms of a clear drop of moderation before and during the election. At the same time it can appear that Instagram has a decrease in the number of SoRs around the election, while Pinterest has an sharp increase right before the election. Figure 2 show in addition that there are great variation for some days regarding the number of SoRs. It would be interesting to analyze these patterns, but this is outside the scope of this paper and the database itself does not provide enough information on the SoRs to test fine grained hypotheses for different explanatory variables. We can see a sharp decrease of SoRs for TikTok during the European Parliament election which can imply that there is a connection between content moderation and the election.

Variables[JHS]

In our analysis we are operating with two main dependent variables. To investigate whether the share of moderation of content categorized as negative effects on civic discourse or elections increases relative to other content the closer to the election (H1), we calculated the share of SoRs of NECDE by dividing the number of SoRs for NECDE on the number of SoRs for all other categories. Because we are operating with small numbers, the variable is created into a percentage for easier interpretation of the results. To investigate whether fully or partially automated moderation of content categorized as NECDE increases relative to other content the closer to the election (H2), we are using the difference between the share of fully or partially automated moderation for the category NECDE subtracted by the share of automated moderation in all categories. The database do not provide or demand any further explanation about how much of the decision was done with human intervention if the SoR is categorized as «partially automated». However, since the moderation decision is automated in some way, it was decided to include it as a automated decision. For easier interpretation the variable is converted into percentage as well. A positive value after the subtraction between the share of automated moderation of NECDE and the share of automated content for the other categories will mean that moderated NECDE content is moderated with some degree of automation at a higher rate compared to all moderated content.

Our main explanatory variable is the distance from the European Parliament election day. Due to our choice of statistical model (linear regression) we specify absolute distance from election day, ranging from 0 to 50, with 0 falling on election day (09.06.2024), as our main explanatory variable, rather than a temporally directed

FIGURE 1 Logarithmic scale of SoR and the category NECDE

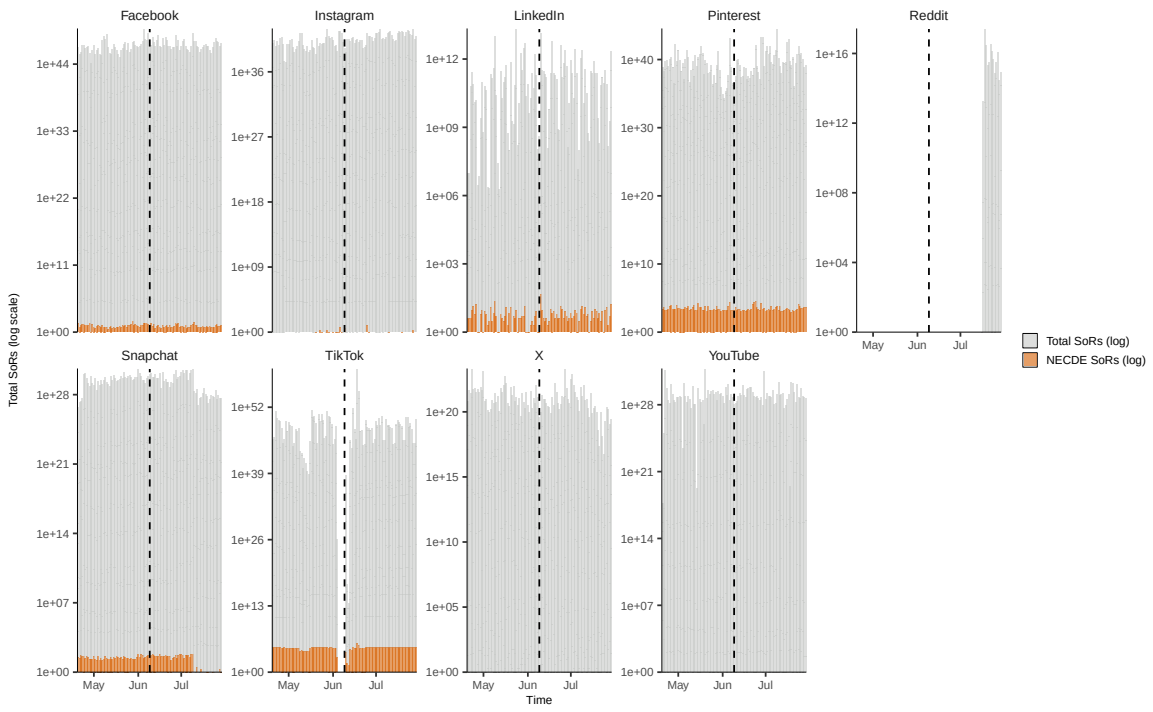
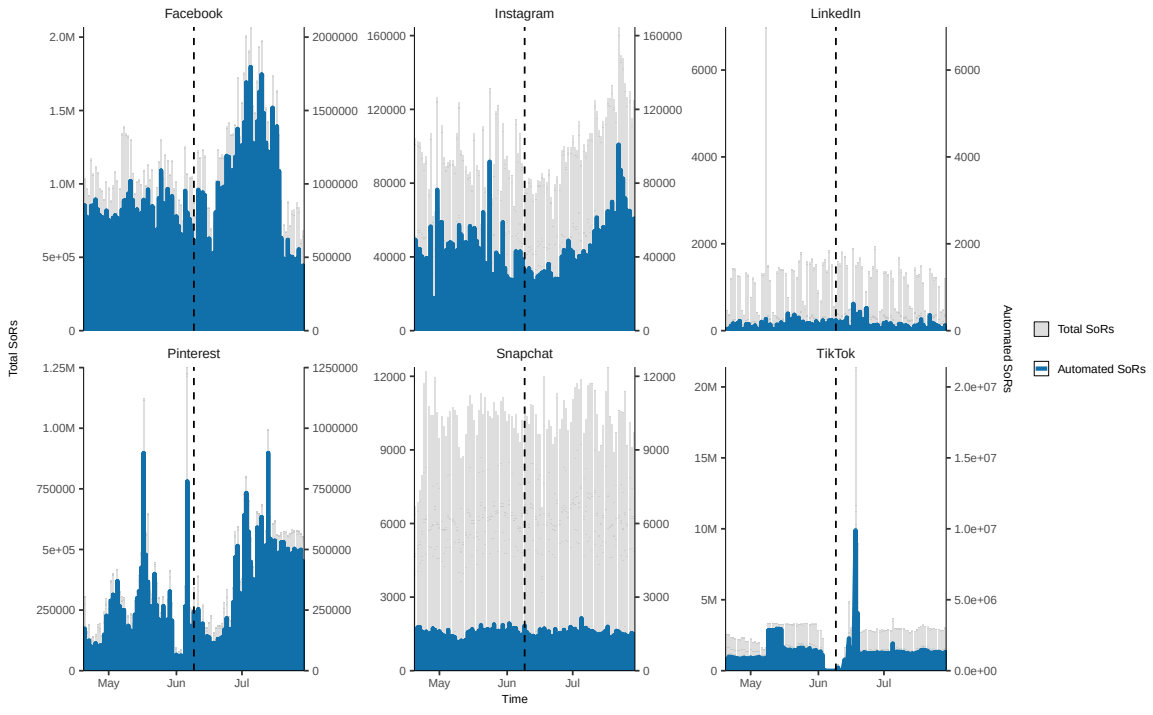


FIGURE 2 Total number of SoRs and automated moderation



counter (i.e. -50 to 50). We are nevertheless interested in a possible difference in trends before and after the election. We discuss below how we proceed to allow for different trends pre- and post-election. We control for if there is a general shift in the moderation trends before or after the election by using a dummy variable where the pre-election period is coded as 1, and as 0 for the post-election period. Our control variable is a weekend dummy as we expect it to be less moderation during the weekend. The variable is coded as 0 if it is a weekday and as 1 if it is a weekend. Finally, we control for the share of SoRs for the category ”Scope of platform service”. The platforms treat the categories differently, where some only report violations as ”Scope of platform service” while other are more varied. Some of the variation in the share of NECDE can be down to this.

Analysis[JHS]

In order to test our hypotheses, we use linear regression models. We first conduct two simple linear OLS regressions for the two dependent variables including the explanatory variables and the control variables. To control if the trend is driven either before or after the election, we include the interaction of pre-election and the election distance. Linear regression models are useful because they provide an intuitive and easily interpretable way to estimate the relationship between explanatory variables and the outcome. However, the social media platforms we are controlling for have a big variation in how much they report on moderation, and how much automated moderation they have, as can be seen in figure 1 and 2. To control for all variables, whether they are observed or not, we are using fixed-effects regression models, including the same variables as for the linear regression. This allows us to control for everything that is constant within that category. In addition to the fixed-effects regression models, we specify a generalized additive model (GAM) to examine the content moderation dynamics around the European election. While the previously models are quite restrictive in their assumptions about the data generating process, GAMs are more flexible, allowing for nonlinear relationships between the predictor and the outcome variable. It is better suited to detect complex or asymmetric temporal patterns, like potential inflection points, plateaus, or rapid changes in moderation behavior that linear models may not capture well.

Placebo testing[JHS]

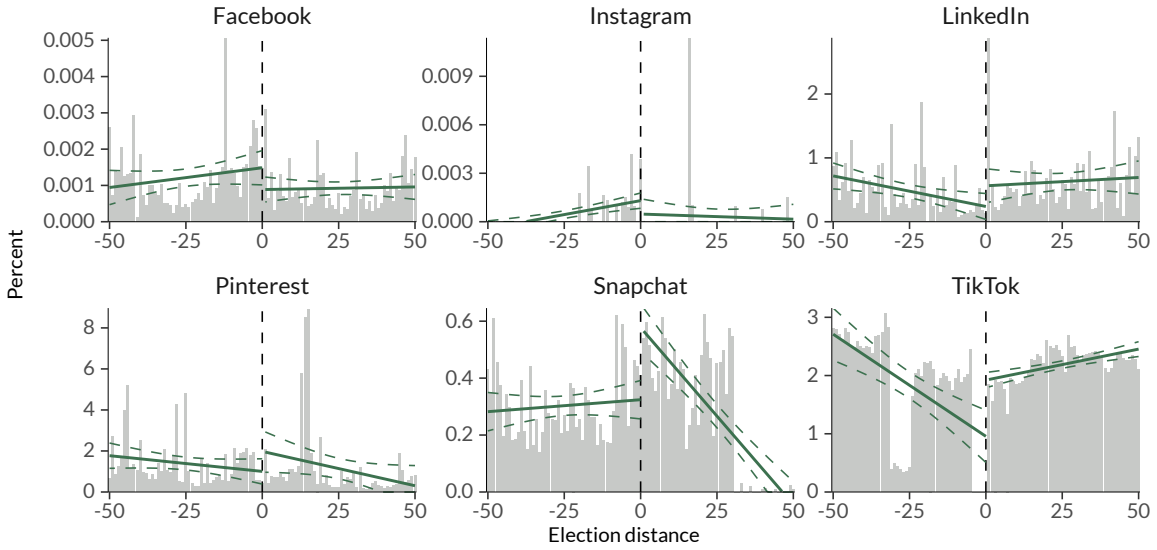
As mentioned previously, the categories for the SoRs are mutually exclusive. This can lead to that we are missing out of relevant information for content in overlapping categories. In addition, we are not certain if the platforms are selective when choosing the SoR category. We are therefore conducting placebo tests for six other categories. We include the categories ”Illegal or harmful speech” and ”Risk for public security” as we expect them to follow the same trend as NECDE. We also include the category ”Scope of platform service” since it is the most reported SoR. Finally, we also include the categories ”Self harm”, ”Unsafe and illegal products” and ”Pornography and sexualized content” which we do not expect to follow the same trend as NECDE for the election time period. We conduct linear OLS regressions including the same independent and control variables as in our main regression. The dependent variables for the placebo tests is the percentage of the share of the specific category (done the same way as for the DV in our main regression for H1).

Results[2]

[2] *Iordanis Pantzartzis*

For descriptives, Figure 3 and Figure 4 plot our outcomes of interest for the 6 platforms we include in our analysis

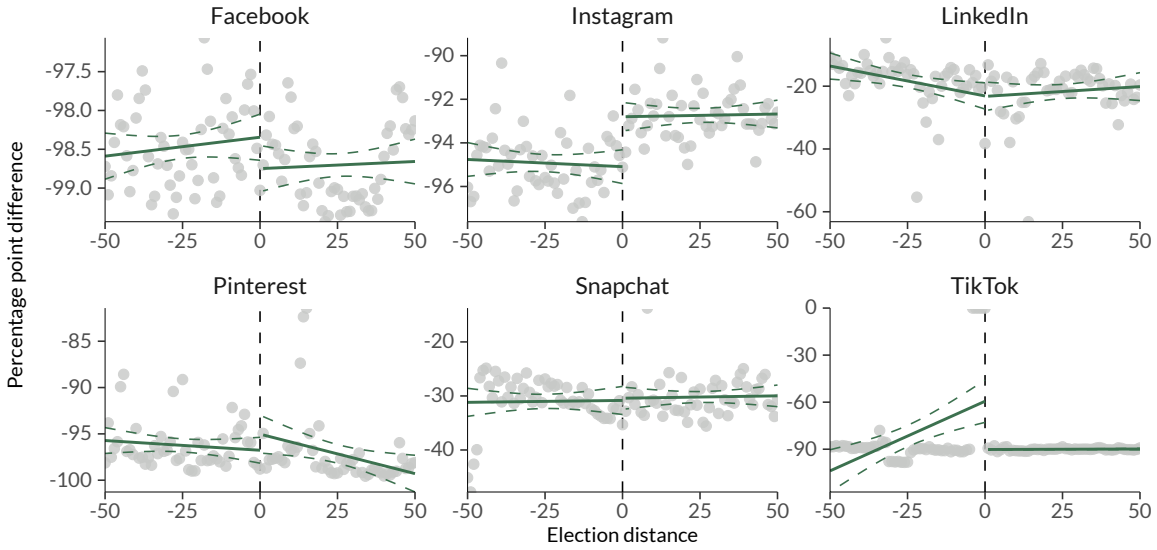
FIGURE 3 Share of SoRs in the NECDE-category (by platform)



against distance from the election. We impose a linear trend before and after the election to obtain a first impression of the relationship between election distance and the respective outcomes of interest. Regarding our first hypothesis (Figure 3), only the trend-lines for Snapchat and Instagram appear to be in line with our expectation, an increase in the share of SoRs in the NECDE-category moving closer to the election. Facebook appears to follow the expected trend in the pre-election period, but not after the election, and vice versa for Pinterest. LinkedIn and TikTok demonstrate a pattern diametrically opposed to our expectation: the share of SoRs in the NECDE-category decreases approaching the election from either direction. Turning to our second hypothesis (Figure 4), first, more generally, all observed values for this difference are negative, so, for all platforms the rate of automated moderation is larger for all SoRs than it is within the NECDE category. Second, none of the platforms conform to our expectation for the difference in rate of automated moderation between NECDE-SoRs and all SoRs, which is that the difference increases moving closer to the election. Facebook and TikTok appear to follow it for the pre-election period, but the trend for TikTok is heavily skewed by a few outliers around and on election day. Pinterest exhibits the expected trend for the post-election period. Instagram and Snapchat appear to show effectively no trend at all and LinkedIn again exhibits trends in both directions that are the opposite to our hypothesis.

Since our hypotheses share the same explanatory variables and only differ in the outcome of interest (H1: share of SoRs in the NECDE-category; H2: share of automated decisions of SoR in the NECDE-category compared to all SoRs), we briefly outline how we proceed for our analytical results for both hypotheses here: First, we employed a directional measure of election distance before now, for illustration purpose, but our main explanatory variable is the *absolute* distance from election day. We perform regressions for four model specifications for each hypothesis. The first specification presents results of a simple linear regression of the outcome of interest on election distance, subsequently adding controls for the weekend and the share of SoRs in the scope-of-platform-service category for the second specification. Next, we add the interaction of election distance and pre-election period to the model specification. This allows for detecting different time trends (positive/negative), or differently steep trends, before and after the election. Finally, we add platform fixed effects to the model specification. The distribution of

FIGURE 4 Share of difference in rate of automated moderation in the NECDE-category (by platform)



SoRs by platform and *category* for the 100 day observation period (see Figure 13 in the appendix) suggests that platform is correlated with category (see Figure 16 in the appendix). A rudimentary calculation of correlation of the independent variables confirms that share of SoRs in the *Scope of platform service* category (henceforth *scope*) is highly correlated with 4 out of the 6 platform-fixed-effects (see Figure 16 in the appendix). As a consequence, we cannot estimate the effect of share of SoRs in the scope category in a fixed effect model. However, we decide to include the share of SoRs in the scope category in the fixed effects model specification for two reasons: First, for the purpose of this paper, we are not primarily interested in the exact between-platform-differences, or the effect of share of SoRs in the scope category directly. Second, excluding the variable from the fixed effects model specification does substantially impact results for our main explanatory variable, election distance (compare Model 4 in Figure 1 and Model 8 in Figure 2 with Figure 15 in the appendix). For our first outcome of interest (share of SoRs in the NECDE category), though only weakly correlated (-0.0203), excluding share of SoRs in the scope category leads the coefficient on election distance to become statistically significant at conventional levels. For our second outcome of interest (difference in automated moderation in the NECDe category vs. all categories), no previously statistically insignificant outcomes now reaches statistical significance, but the coefficient magnitudes do differ noticeably.

To preempt our results, most of our findings are nulls, nevertheless (or rather because of this), we will sketch out generalized interpretations for possible estimate values here, i.e. we discuss beforehand the implication the sign on the coefficients (positive or negative) would entail, as they are somewhat counter-intuitive. First, in both our hypotheses we state the expectation to observe a *positive* trend in the outcome of interest *as election day approaches*. The independent variable is, however, implemented as the *absolute temporal distance* from the election, therefore a positive trend in the outcome approaching election day would manifest as a *negative coefficient* on the trend estimate. Put in terms of election distance, with increasing temporal distance from election day, we expect a *negative* trend in the outcome of interest. Second, as a consequence of including an interaction term, the interpretation of our main variable of interest, election distance, changes from the first two model specifications to the latter two. The latter two specifications interact election distance with a dummy for the pre-election period, therefore the coefficient

TABLE 1 OLS regression percentage of SoRs in the NECDE-category on election distance

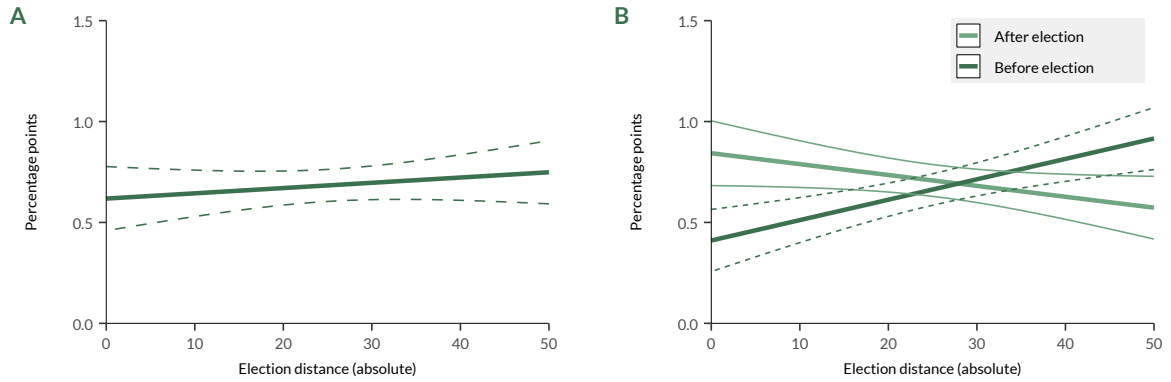
	Model 1		Model 2		Model 3		Model 4	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Constant	0.622***	(0.082)	0.813***	(0.093)	1.035***	(0.124)		
Explanatory variables								
Election distance	0.002	(0.003)	0.003	(0.003)	−0.005	(0.004)	−0.005+	(0.003)
Pre-election					−0.432**	(0.161)	−0.433***	(0.113)
El. dist.*Pre-election					0.016**	(0.006)	0.016***	(0.004)
Controls								
Weekend			0.028	(0.089)	0.041	(0.088)	0.041	(0.062)
% SoRs scope-category			−0.703***	(0.139)	−0.705***	(0.139)	−0.759*	(0.317)
Num.Obs.	606		606		606		606	
R2	0.001		0.042		0.055		0.683	
R2 Adj.	0.000		0.037		0.047		0.677	

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

on election distance is interpreted as the predicted trend for the outcome on interest *after the election*, i.e. where *pre-election* = 0. The predicted trend for the pre-election period is computed as a combination of the coefficient on election distance and the coefficient on the interaction ($\beta_{\text{election distance}} + \beta_{electiondistance*pre-election} * 1$). Again, to support our hypotheses, we would expect both of these estimates in the latter two model specifications to be *negative* as election distance increases, indicating a positive trend, as election day approaches. Finally, the coefficient on the pre-election dummy is the intercept change moving from the post-election period (0) to the pre-election period (1) on election day (*election distance* = 0). Put graphically, the coefficient on the pre-election dummy indicates the y-shift in from the beginning of the post-election trend-line to the end of the pre-election trend-line (which meet on election day), associated with the dummy variable. The coefficient on the interaction term itself indicates the difference in the slope of the pre-election trend lines compared to the post-election trend line. However, as previously described, because we implement our explanatory variable as *absolute* election distance and the ensuing reversal in direction for the pre-election period, we cannot rely on this estimate to establish whether there is a statistically significant difference between the slopes in the trends in the forward direction of time. Since, the sign on the pre-election trend will necessarily be inverted in our model compared to a chronologically ordered counter, the difference in point estimates of the slopes will be artificially larger or smaller (depending on the specific results) compared to a chronologically ordered counter. Table 12 in the appendix illustrates this point for hypothesis 1 (percentage of SoRs in the NECDE-category, data from model 4, see Table 1).

Table 1 reports results for Hypothesis 1 (percentage of SoRs in the NECDe-category). The coefficient on election distance is not statistically significant at conventional levels for any model specification. Only when platform fixed effects are included in the model (i.e. model 4), does the estimate approach conventional levels of statistical significance ($p = 0.0531$). Though not statistically significant, the signs on the coefficients on election distance in models 1 and 2 are *positive*, the opposite direction of our prediction. A positive trend of the percentage of SoRs in the NECDE-category, as election distance increases, all else held constant, would suggest a predicted decrease *moving towards election day*. For illustration of the reversal in direction for the pre-election period, see Figure 5, which plots the predicted trends against election distance and Figure 6, which plots the fitted trends chronologically. In model specifications 3 and 4, the signs on the coefficients on election distance do point in the theorized direction, though,

FIGURE 5 Predicted percentage of SoRs in the NECDE-category (election distance)

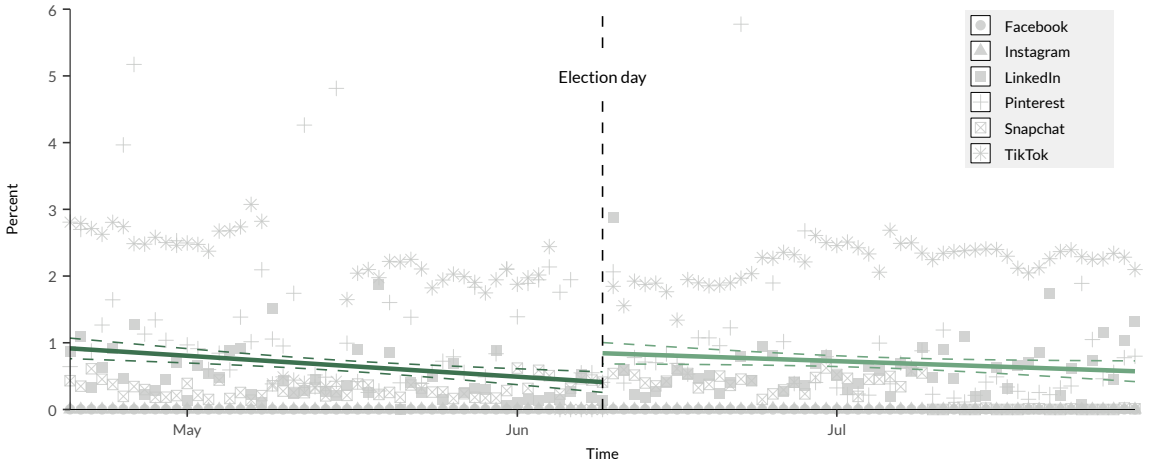


Note: **Panel A**: Predicted trend based on model 2 results. **Panel B**: Predicted trend based on model 4 results (incl. interaction effect, platform-fixed effects). Dashed lines show 95% confidence interval.

again, they are not, or only marginally, statistically significant (Estimating model 4 without the share of SoRs in the scope category, the coefficient on election distance does reach statistically significance at the 5% level,see Table 15). Due to the interaction, the coefficient on election distance in models 3 and 4 indicates the trend for the share of the SoRs in the NECDE-category after the election. The trend before the election is given by the combined terms of election distance and the interaction term. Figure 7A and Table 7B report the partial marginal trend for election distance on the percentage of SoRs in the NECDE-category, conditional on the time period (before vs. after the election) using results from model 4, which includes platform fixed effects. The point estimate for the pre-election trend (0.0101) is statistically different from zero $(p < 0.001, \text{SE} = 0.00270)$, but in the opposite direction to our expectation. All else held constant, as election distance increases in the pre-election period, the model predicts a statistically significant increase in the share of SoRs in the NECDE-category of 0.0101 percentage points per day, or, conversely, a decrease of 0.0101 percentage points per day, *moving towards the election*, refuting Hypothesis 1. The coefficient on the pre-election dummy is negative and highly statistically significant in both models 3 and 4. The sign indicates there is a downward y-shift moving from the post-election trend to the pre-election trend on election day in the predicted percentage of SoRs in the NECDE-category, or in the forward direction of time, a *jump* moving from the pre-election trend to the post-election trend, all else held constant (see the y-intercept of the two lines in the B panel of Figure 5 and Figure 6).

Table 2 reports results for Hypothesis 2. Again the coefficient on election distance fails to reach statistical significance at conventional levels in any of the model specifications. The signs on the election distance estimates in models 5 and 6 (no interaction effect) would be in line with our hypothesis. In model specifications 7 and 8, which add the interaction term, the sign on election distance flips, which would have indicated a trend in the opposite direction of our hypothesis for the post-election period, i.e. the rate of automated moderation of SoRs in the NECDE-category would be predicted to increase as distance from the election increases, or, decrease, moving towards the election. Again, for illustration purpose, see Figure 8 (plotted on election distance) and Figure 9 (chronological). The coefficient on the pre-election dummy is statistically significant and positive in model 8, this time indicating a *drop* in the predicted difference in automated moderation on election day, moving from the pre-election period to the post-election period (see y-intercepts of the lines in panel B of Figure 8 and Figure 9). The predicted trend for the pre-election period is computed as before and results are shown in Figure 10A and Table 10B.

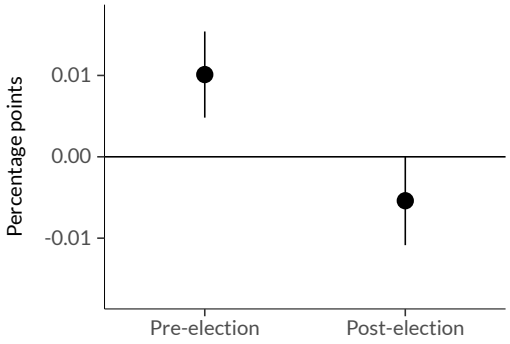
FIGURE 6 Predicted percentage of SoRs in the NECDE-category (chronological order)



Note: Y-axis truncated at 6 for readability (3 data points not shown). Dashed lines show 95% confidence interval.

FIGURE 7 Slope estimate, predicted percentage of SoRs in the NECDE-category, pre- and post-election

A



B

	Est.	95% CI
Post-election	-0.0054+	[-0.0109, 0.0001]
Pre-election	0.0101***	[0.0048, 0.0154]

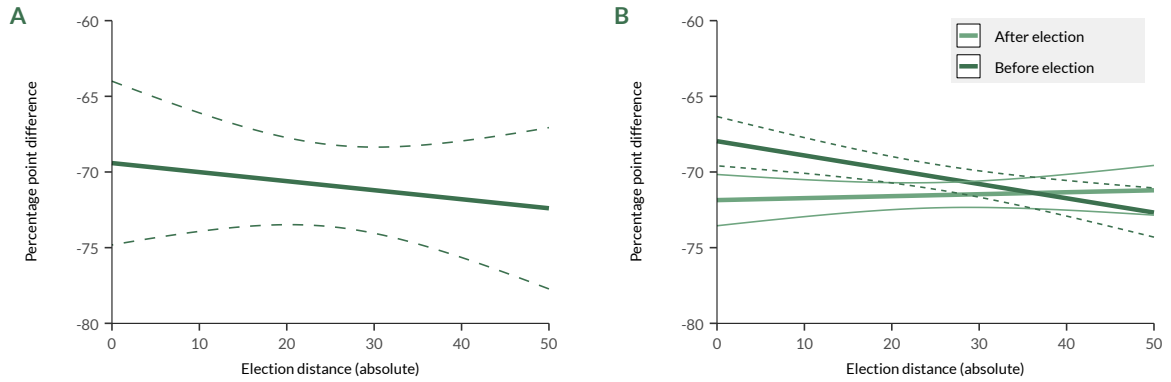
Note: + p <0.1, * p <0.05, ** p <0.01, *** p <0.001. Data from model 4 (includes platform fixed effects).

TABLE 2 OLS regression percentage of SoRs in the NECDE-category on election distance

	Model 5		Model 6		Model 7		Model 8	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Constant	−69.411***	(2.755)	−50.608***	(2.618)	−52.553***	(3.522)		
Explanatory variables								
Election distance	−0.060	(0.094)	−0.041	(0.078)	0.015	(0.113)	0.013	(0.029)
Pre-election					3.779	(4.575)	3.894**	(1.196)
El. dist.*Pre-election					−0.106	(0.157)	−0.107**	(0.041)
Controls								
Weekend			−0.725	(2.497)	−0.844	(2.504)	−0.905	(0.654)
% SoRs scope-category			−65.800***	(3.933)	−65.762***	(3.938)	−56.740***	(3.351)
Num.Obs.	606		606		606		606	
R2	0.001		0.318		0.319		0.991	
R2 Adj.	−0.001		0.315		0.313		0.991	

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

FIGURE 8 Difference in automated moderation (election distance)

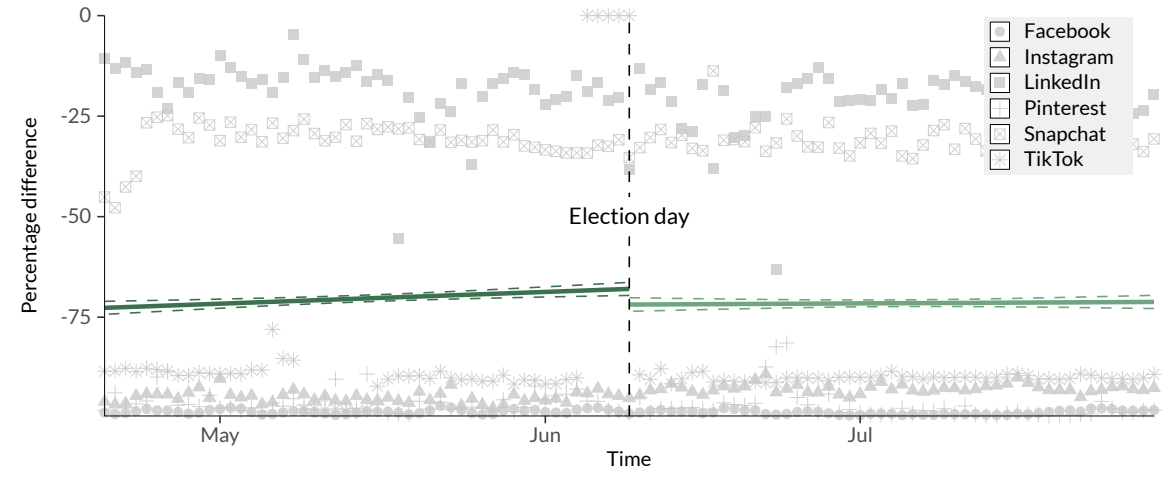


Note: **Panel A**: Predicted trend based on model 6 results. **Panel B**: Predicted trend based on model 8 results (incl. interaction effect, platform-fixed effects). Dashed lines show 95% confidence interval.

At −0.0943, the fitted trend for the pre-election period is statistically different from zero ($p < 0.001$): As election distance increases, the predicted difference in rate of automated moderation of SoRs in the NECDE-category decreases by −0.0943 per day. Or, as election day moves closer, the predicted rate of automated moderation of SoRs in the NECDE-category *increases* by 0.0943 percentage points every day compared to the rate of automated moderation in all categories, which is evidence in favor of our hypothesis.

As hinted at in previous sections, the DSA transparency database itself has some major limitations, resulting from its structure and data generating process, which potentially limit the internal validity of our results. First and foremost, we have good reason to believe that platforms, who report their content moderation decisions to the database, do not treat the category variable the same. Figure 13 (see Appendix) plots the the share of SoRs in each category for the entire period of observation for each platform. Facebook, Instagram, LinkedIn and TikTok categorize their SoRs mostly as "Scope of platforms service" and don't make much use of the other categories. This suggests that platform do not take great care to categorize SoRs appropriately within the 13 category framework envisioned by the DSA, and we cannot entirely trust that the NECDE-category actually captures the moderation

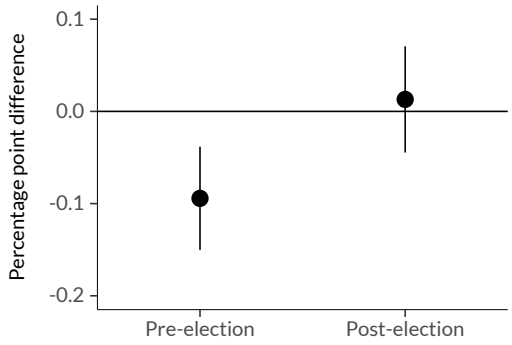
FIGURE 9 Difference in automated moderation (chronological order)



Note: Dashed lines show 95% confidence interval.

FIGURE 10 Slope estimate, predicted difference in automated moderation, pre- and post-election

A Something



B Something

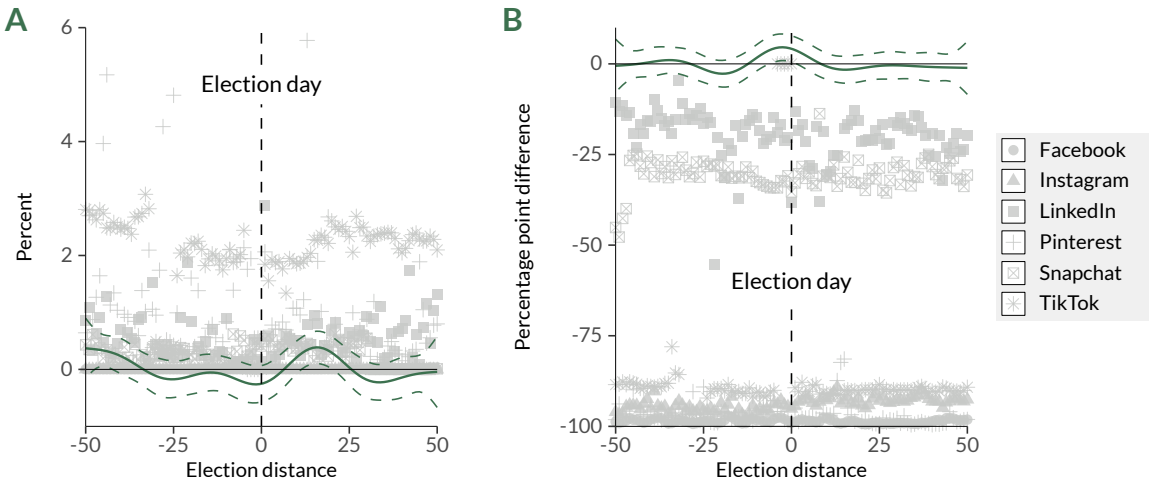
	Est.	95% CI
Post-election	0.0130	[-0.0447, 0.0707]
Pre-election	-0.0943***	[-0.1502, -0.0384]

Note: + p <0.1, * p <0.05, ** p <0.01, *** p <0.001. Data from model 8 (includes platform fixed effects).

dynamics of misinformation. This is compounded by the issue that the NECDe category on its face already is not a perfect proxy for mis- and disinformation. Unlike the issue in the data generating process, we can address the latter by investigating the dynamics of moderation in other categories. Our theoretical framework does not necessarily immediately suggest any strong and clear expectations for contend moderation dynamics in the other categories, nevertheless, operating from semantic meaning of the category name alone, one the one hand, the categories "Illegal or harmful speech" and "Risk for public security" can reasonably be expected to move in the same direction as our hypotheses suggest for the NECDE category, while we do not expect "Self harm", "Scope of platform services", "Unsafe and illegal products" and "Pornography and sexualized content" to show any systematic relationship with the incidence of the European election. We conduct this robustness check by rerunning our main regression, varying the outcome of interest. Specifically we rerun our fourth regression specification (full set of controls, interaction effect and platform fixed effects). See Table 17 in the appendix for the full set of results. We do find statistically significant correlations with election distance for "Risk to public security", "Pornography and sexualized content" and "Self harm" for the period after the election. The sign on "Risk to public security", for which we expected to find an association is negative, i.e. the share of SoRs in that category decreases with increasing distance to the election, or increases moving towards the election, so, in line with our hypothesis for the NECDE-category. For the pre-election period we only find a statistically significant effect in the expected direction (increasing as election day comes closer) (see Figure 14 in the appendix).

Finally, we run nonparametric regressions of our main outcomes of interest (percentage of SoRs in the NECDE-category, difference in rate of automated moderation of NECDE-SoRs and all others). This alleviates some of the constrains of linear models, such as capturing non-linear trends, and more germane to our case, allowing us to fit a model on our explanatory variable directly, i.e. election distance ranges from $-50 : 50$ rather than being the absolute value. We fit a GAM regression of the outcomes of interest on election distance as the smooth and include controls for the weekend and platform fixed effects (see Figure 11 for results). For both our hypotheses the fitted curves produced by the nonparametric regression show some signs of non-linearity, but their patterns are overall not distinct enough cto produce a statistically significant result. This results is in line with the findings from our linear models, before adding interactions and platform fixed effects.

FIGURE 11 Nonparametric regression



Note: Dashed lines show 95% confidence interval.

Conclusion[1]

[1] *David Breukel*

Many people turn towards social media to inform themselves about politics – and the importance of online platforms as a news source is likely to grow in the coming years. At the same time, the quality of information and civic discourse on social media platforms is frequently called into question. In recent years, the suspicion that important political events are influenced by misinformation and foreign interference on social media has become widespread. Platforms are, however, not without means to counter content with negative effects on civic discourse or elections. They can block or otherwise moderate online content at a large scale; But content moderation is a non-transparent tool.

In this contribution, we study moderation dynamics on six social media platforms around the 2024 European Parliament election. Relying on the DSA Transparency Database, we analyze 100 days of content moderation by examining the relative frequency of moderation decisions targeting content that is harmful to civic discourse or elections, as well as the extent to which these decisions are automated compared to decisions over other content types.

Our results contradict our theoretical expectations. Our main model shows that, as the election approaches, the relative frequency of moderation decisions targeting content classified as harmful to civic discourse decreases. Immediately after the election, there is an increase in this share. As time goes on, the share decreases again – although this estimate is not statistically significant. This suggests that increases in online misinformation on social media (Ferrara, 2017; Grinberg et al., 2019a) may not necessarily be met with a corresponding increase in content moderation. Instead, the decline in the relative frequency of moderation may suggest that platforms are more hesitant to moderate when stakes are high. In addition, our results suggest that the share of automated moderation decisions concerning content deemed harmful to civic discourse decreases compared to the share of automated decisions over other content prior to election. This pattern raises important questions about whether automated moderation effectively addresses election-related harms. However, these conclusions are complicated by the results of placebo tests. When we re-estimate our models using content categories that are supposedly unrelated to elections – such as

pornography and self-harm –, the observed effects remain of comparable magnitude. This casts doubt on the link between the observed moderation patterns and the electoral context.

Our study remains subject to limitations that future research should address. First, we offer results from the study of one case, which raises questions about the generalizability of our findings. While European Parliament elections ask an extraordinary amount of individuals to the ballot box, they are frequently considered to be second-order elections (e.g., Hobolt & Wittrock, 2011; Reif & Schmitt, 1980). Such elections are expected to have lower turnout, to be dominated by national issues, and to penalize parties in national government (Reif & Schmitt, 1980). More generally, stakes are perceived to be lower in second-order elections. In this case, should we expect less moderation related to content with negative effects on public discourse? On the one hand, it is possible that incentives to spread disinformation or to manipulate public opinion are lower in second-order elections. Additionally, platforms may channel their moderation resources towards high-stakes elections, where harmful content could be more detrimental. On the other hand, harmful content during second-order elections could still erode public trust in democracy, persisting over time and thus influencing first-order elections later on. Future research should strive to test if our results travel to other elections. Unfortunately, the DSA does not include the origin of moderated content. While platforms can report the language of moderated content, they very rarely do so. This poses a challenge to studying content moderation dynamics in response to political events within specific countries.

A second limitation pertains to our study's lack of knowledge about the actual content. The DSA does not require platforms to disclose the actual piece of content that they chose to moderate. This is especially problematic given that platforms seem to interpret the categories within the DSA quite differently, leading to many anomalies within the Transparency Database. Some social media platforms – like Youtube and X – never classify content as harmful to civic discourse or elections. Anecdotal and empirical evidence suggest that this is unlikely to be the case because no such content exists on these platforms (Guess & Lyons, 2020; S. Thompson, 2024). A possible explanation for the different patterns is that platforms use the category relating to the scope of platform conditions to different extents. Especially for X, where moderation staff has been vastly reduced (Laaff, 2024), the scope of platform conditions may serve as a catch-all category. One way for researchers to gain a more detailed understanding of content moderation dynamics around elections is to collaborate with social media platforms to obtain data on the content as well as moderation decisions. This would allow them to judge the reliability of classification decisions made by human moderators and algorithms. Prior research suggests that both human and algorithmic moderation positively affect user behavior (Horta Ribeiro et al., 2023; Srinivasan et al., 2019), but it remains unclear if these moderation types differ in their quality.

Such a collaboration would also empower future research to address a related, third limitation of the study. In this essay, we rely on the share of moderation decisions that relates to content with negative effects on civic discourse to test our first hypothesis. A more direct test of our research question may compare the proportion of moderated content to the total volume of content exhibiting such effects. This would allow researchers to examine trends in misinformation and similar issues around elections, as well as platforms' responses to them in an unprecedented manner.

These limitations notwithstanding, our study makes some important contributions to the literature on content moderation. First, we provide initial evidence on how elections affect content moderation dynamics, both in terms of the prevalence of decisions on content that may be politically harmful and the extent to which they are

automated. In addition, the study provides broad coverage, including six major social media platforms and the entire European Union, whereas prior research has generally focused on single platforms in the US. Lastly, we highlight some of the advantages and drawbacks of the DSA Transparency Database for answering our and similar research questions.

References

Bovet, A., & Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, *10*(1), 7.

Commission, E. (2022). Article 24, transparency reporting obligations for providers of online platforms - the digital services act (dsa) [Accessed: 2025-03-19]. https://www.eu-digital-services-act.com/Digital_Services_Act_Article_24.html

Commission, E. (2025a). Digital services act [Accessed: 2025-03-19]. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Commission, E. (2025b). Digital services act transparency portal [Accessed: 2025-03-19]. https://transparency.dsa.ec.europa.eu

European Commission. (2024a). Dsa transparency database - questions and answers [Accessed: 2024-03-18]. https://digital-strategy.ec.europa.eu/en/faqs/dsa-transparency-database-questions-and-answers

European Commission. (2024b). *DSA Transparency Database: Documentation*. Retrieved March 10, 2025, from https://transparency.dsa.ec.europa.eu/page/documentation

European Commission. (2025). Digital services act transparency database - documentation [Accessed on March 18, 2025]. https://transparency.dsa.ec.europa.eu/page/documentation

European Parliament. (2024a). *Flash Eurobarometer 3153: Media and news survey 2023* (Version 1.0.0). GESIS. https://doi.org/10.4232/1.14244

European Parliament. (2024b). *Metsola at the European Council: This election will be the test of our systems*. Retrieved March 10, 2025, from https://www.europarl.europa.eu/news/de/press-room/20240321IPR19532/metsola-at-the-european-council-this-election-will-be-the-test-of-our-systems

Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*. https://doi.org/10.5210/fm.v22i8.8005

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

González-Bailón, S., d'Andrea, V., Freelon, D., & Domenico, M. D. (2023). The advantage of the right in social media news sharing. *Nature Human Behaviour*, *7*, 1081–1091. https://doi.org/https://doi.org/10.1093/pnasnexus/pgac137

González-Bailón, S., Lazer, D., Barberá, P., Godel, W., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., et al. (2024). The diffusion and reach of (mis) information on facebook during the us 2020 election. *Sociological Science*, *11*, 1124–1146.

Greene, C., Nash, R., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory*, *29*(5), 587–604. https://doi.org/10.1080/09658211.2021.1923754

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019a). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019b). Fake news on twitter during the 2016 u.s. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy* (pp. 10–33). Cambridge University Press.

Guriev, S., Melnikov, N., & Zhuravskaya, E. (2021). 3g internet and confidence in government. *The Quarterly Journal of Economics*, *136*(4), 2533–2613.

Herszenhorn, D. M., & Volpicelli, G. (2024). How russia is targeting the European election. *Politico*. https://www.politico.eu/newsletter/eu-election-playbook/how-russia-is-targeting-the-european-election/

Hobolt, S. B., & Wittrock, J. (2011). The second-order election model revisited: An experimental test of vote choices in European Parliament elections. *Electoral Studies*, *30*(1), 29–40. https://doi.org/https://doi.org/10.1016/j.electstud.2010.09.020

Hong, S. (2013). Who benefits from twitter? social media and political competition in the u.s. house of representatives. *Government Information Quarterly*, *30*(4), 464–472. https://doi.org/https://doi.org/10.1016/j.giq.2013.05.009

Horta Ribeiro, M., Cheng, J., & West, R. (2023). Automated content moderation increases adherence to community guidelines. *Proceedings of the ACM Web Conference 2023*, 2666–2676. https://doi.org/10.1145/3543507.3583275

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on twitter. *Proceedings of the national academy of sciences*, *119*(1), e2025334119.

Ivaldi, G., & Zankina, E. (2024). *2024 ep elections under the shadow of rising populism*. European Center for Populism Studies (ECPS).

Kang, C., & Satariano, A. (2024). Social media companies face global tug-of-war over free speech. *The New York Times*. Retrieved February 25, 2025, from https://www.nytimes.com/2024/12/30/technology/trump-administration-speech-policy.html

Kottasová, I. (2024). Doppelgängers and deepfakes: How Russian trolls are meddling in the world's second-biggest democratic vote. *CNN*. Retrieved February 25, 2025, from https://www.cnn.com/2024/06/04/climate/russia-disinformation-eu-elections-intl/index.html

Laaff, M. (2024). Social media: X lässt das mal so stehen. *Die Zeit*. Retrieved February 25, 2025, from https://www.zeit.de/digital/internet/2024-05/x-online-plattform-eu-kommission-content-moderation

Munger, K., Egan, P. J., Nagler, J., Ronen, J., & Tucker, J. (2022). Political knowledge and misinformation in the era of social media: Evidence from the 2015 uk election. *British Journal of Political Science*, *52*(1), 107–127.

Pierri, F., Luceri, L., Chen, E., & 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), 43.

Reif, K., & Schmitt, H. (1980). Nine second-order national elections – a conceptual framework for the analysis of European election results. *European Journal of Political Research*, *8*(1), 3–44. https://doi.org/10.1111/j.1475-6765.1980.tb00737.x

Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. *Journal of Information Technology*, *36*(4), 409–426.

Schaub, M., & Morisi, D. (2020). Voter mobilisation in the echo chamber: Broadband internet and the rise of populism in europe. *European Journal of Political Research*, *59*(4), 752–773.

Soave, R. (2022). Bari Weiss Twitter Files reveal systematic 'blacklisting' of disfavored content. *Reason*. Retrieved February 25, 2025, from https://reason.com/2022/12/09/bari-weiss-twitter-files-elon-musk-blacklist-shadow-banning/

Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 163:1–163:21. https://doi.org/10.1145/3359265

Thompson, A. (2020). Why the right wing has a massive advantage on Facebook. *Politico*. Retrieved February 25, 2025, from https://www.politico.com/news/2020/09/26/facebook-conservatives-2020-421146

Thompson, S. (2024). 5 days with Elon Musk on X: Deepfakes, falsehoods and lots of memes. *The New York Times*. Retrieved February 25, 2025, from https://www.nytimes.com/2024/09/27/technology/elon-musk-x-posts.html

Trujillo, A., Fagni, T., & Cresci, S. (2023). The dsa transparency database: Auditing self-reported moderation actions by social media. *arXiv preprint arXiv:2312.10269*.

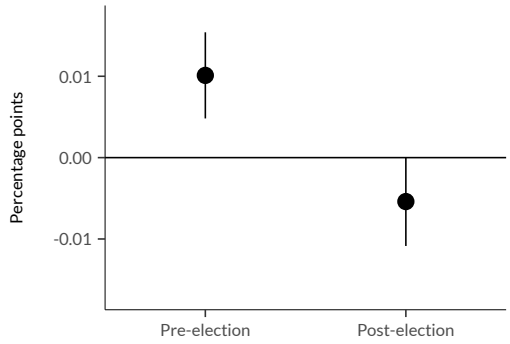
Waisbord, S. (2018). The elective affinity between post-truth communication and populist politics. *Communication Research and Practice*, *4*(1), 17–34. https://doi.org/10.1080/22041451.2018.1428928

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, *12*, 415–438. https://doi.org/https://doi.org/10.1146/annurev-economics-081919-050239

Appendix

FIGURE 12 Trend-line slope estimates (absolute vs. chronological distance)

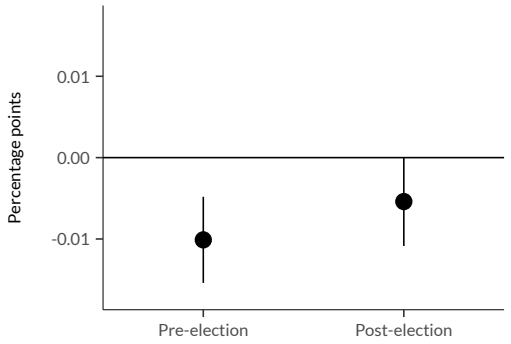
A Election distance (absolute)



	Est.	95% CI
Post-election	-0.0054+	[-0.0109, 0.0001]
Pre-election	0.0101***	[0.0048, 0.0154]

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

B Election distance (chronological)



	Est.	95% CI
Post-election	-0.0054+	[-0.0109, 0.0001]
Pre-election	-0.0101***	[-0.0154, -0.0048]

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

FIGURE 13 Share of difference SoR categories over 100 day observation period (by platform)

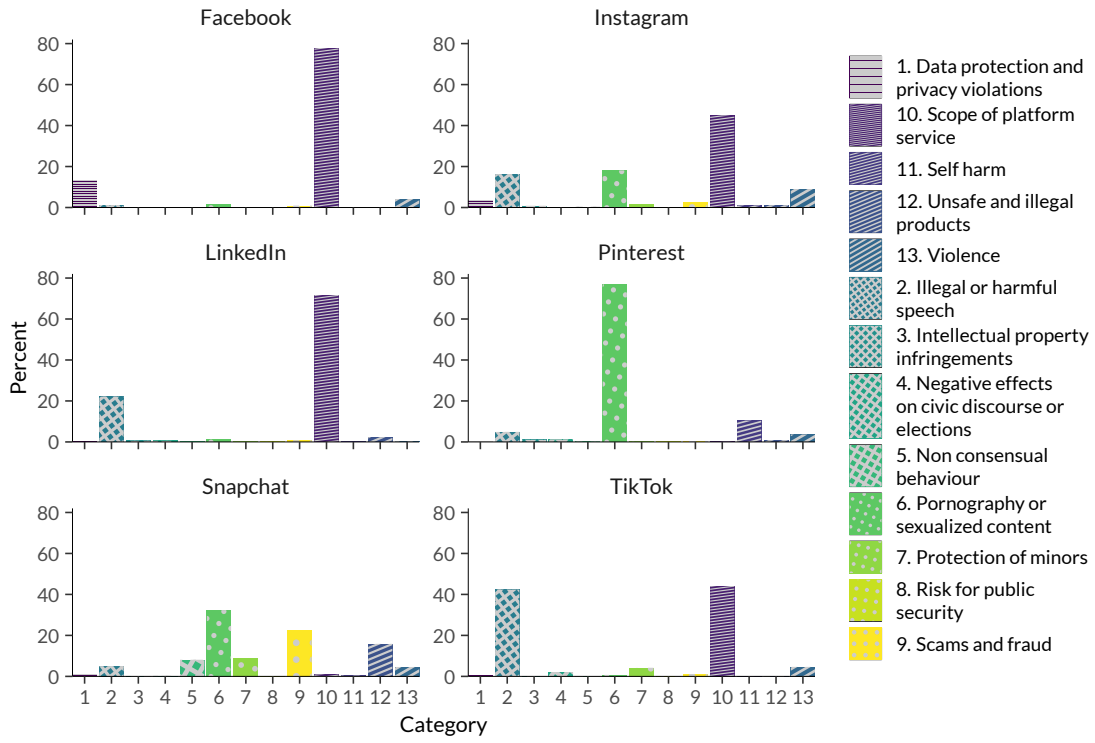


FIGURE 14 Trend-line slope estimates pre- and post election for different outcome variables

		Election distance	
		Est.	95% CI
Illegal speech	Post-election	−0.028	[−0.073, 0.016]
	Pre-election	0.036	[−0.007, 0.079]
Illegal products	Post-election	0.000	[−0.009, 0.009]
	Pre-election	0.005	[−0.004, 0.013]
Scope of platform service	Post-election	0.018	[−0.053, 0.088]
	Pre-election	0.034	[−0.034, 0.102]
Public security	Post-election	−0.001***	[−0.001, 0.000]
	Pre-election	−0.001**	[−0.001, 0.000]
Pornography	Post-election	0.062**	[0.025, 0.100]
	Pre-election	−0.024	[−0.060, 0.013]
Self-harm	Post-election	−0.044**	[−0.070, −0.017]
	Pre-election	0.010	[−0.016, 0.036]

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

FIGURE 15 OLS regression results, fixed effects models without scope category control

A

	Model 9 (H1)	Model 10 (H2)
Election distance	−0.006*	0.003
	(0.003)	(0.036)
Pre-election	−0.423***	4.611**
	(0.114)	(1.454)
El. dist*Pre-election	0.015***	−0.117*
	(0.004)	(0.050)
Weekend	0.036	−1.290
	(0.062)	(0.796)
Num.Obs.	606	606
R2	0.680	0.987
R2 Adj.	0.674	0.987

B

	Model 9 (H1)	Model 10 (H2)
Post-election	-0.006*	0.003
	[-0.011, -0.000]	[-0.067, 0.073]
Pre-election	0.010***	-0.114**
	[0.005, 0.015]	[-0.182, -0.046]

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

FIGURE 16 Correlation of the independent variables (models 4 and 8)

Variable	Election distance	Pre-election	Weekend	Share scope-category	Facebook	Instagram	LinkedIn	Pinterest	Snapchat	TikTok	Interaction
Election distance	1	0.627	-0.059	-0.02	-0.245	-0.388	-0.617	-0.674	-0.674	-0.416	-0.718
Pre-election	0.627	1	-0.057	0.035	-0.251	-0.363	-0.53	-0.557	-0.557	-0.384	-0.866
Weekend	-0.059	-0.057	1	-0.035	-0.017	-0.048	-0.104	-0.126	-0.126	-0.054	0.043
Share scope-category	-0.02	0.035	-0.035	1	-0.92	-0.8	-0.373	0.005	0.006	-0.768	-0.013
Facebook	-0.245	-0.251	-0.017	-0.92	1	0.867	0.545	0.214	0.213	0.846	0.202
Instagram	-0.388	-0.363	-0.048	-0.8	0.867	1	0.608	0.331	0.33	0.828	0.301
LinkedIn	-0.617	-0.53	-0.104	-0.373	0.545	0.608	1	0.515	0.515	0.617	0.454
Pinterest	-0.674	-0.557	-0.126	0.005	0.214	0.331	0.515	1	0.557	0.353	0.484
Snapchat	-0.674	-0.557	-0.126	0.006	0.213	0.33	0.515	0.557	1	0.352	0.484
TikTok	*-0.416*	*-0.384*	*-0.054*	*-0.768*	*0.846*	*0.828*	*0.617*	*0.353*	*0.352*	*1*	*0.32*
Interaction	-0.718	-0.866	0.043	-0.013	0.202	0.301	0.454	0.484	0.484	0.32	1

FIGURE 17 OLS regression of share of different SoR categories on election distance

	Illegal speech	Illegal products	Scope of platform service	Public security	Pornography	Self-harm
Constant	13.822***	0.444	76.860***	0.060**	6.375***	1.232
	(2.146)	(0.432)	(1.341)	(0.019)	(1.823)	(1.287)
Explanatory variables						
Election distance	−0.028	0.000	0.018	−0.001***	0.062**	−0.044**
	(0.023)	(0.005)	(0.036)	(0.000)	(0.019)	(0.014)
Pre-election	−3.594***	0.084	−1.265	0.010	1.272	−1.020+
	(0.917)	(0.184)	(1.461)	(0.008)	(0.779)	(0.550)
Election distance * Pre-election	0.064*	0.005	0.016	0.000	−0.086**	0.054**
	(0.031)	(0.006)	(0.050)	(0.000)	(0.027)	(0.019)
Controls						
Weekend	0.144	0.109	0.679	0.000	0.062	0.205
	(0.502)	(0.101)	(0.800)	(0.004)	(0.426)	(0.301)
Share of SoRs in scope-category	-14.542***	−0.473		−0.035	−7.451***	−0.345
	(2.569)	(0.517)		(0.023)	(2.182)	(1.541)
Num.Obs.	606	606	606	606	606	606
R2	0.860	0.681	0.906	0.584	0.970	0.570
R2 Adj.	0.858	0.676	0.904	0.577	0.969	0.563

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Resulst for "Scope of platform service" do not include share of SoRs in the tha category as control.