Auditing Content Moderation on Social Media Platforms

Edited by Richard Rogers

Digital Methods Initiative, Humanities Labs,
University of Amsterdam

March 2026

Table of contents

How to report this content? Auditing content moderation on social media platforms

Richard Rogers

This collection of work concerns the outcomes of reporting content to social media platforms in violation of platform rules. Are there systems in place for users to report rule-offending content? How do platform systems respond, both to initial reports as well as to appeals when available? What can we learn more generally about how platforms (automatically and manually) moderate violating content reported to them?

To answer these questions, in the following we introduce 'moderation auditing' (Sekwenz et al., 2025), that is, the study of reporting rule-offending content to platforms through a series of test cases. Researchers identify what we call 'reportable content' on the platforms and consider why it should be reported in the first place and how to make it known to them. We also describe the labour involved in such reporting and the prospects of user-assisted content moderation more generally (Malhotra, 2023).

The context is the run-up to national elections in Europe when there is heightened scrutiny of the availability of social media content that may violate their integrity (Vliegenthart et al., 2024). Posts that impinge upon election integrity interfere with electoral proceedings (Chiarella, 2021); it can be as elementary as false information about voting such as the date of the elections and how and where to vote (Udupa, 2019).

But it often refers to disinformation campaigns and influence operations conducted by malicious (external) actors that manufacture attention through such practices as algorithmic manipulation and coordinated inauthentic behaviour (Bradshaw & Howard, 2018; Hiltunen, 2021). A standard term is foreign information manipulations and interference, or FIMI (Proto et al. 2025), though in practice it may be fused (wittingly or unwittingly) with domestic information operations as well as content creation such as meme seeding that adds to the campaigns but does so for (commercial) purposes such as engagement farming.

The studies presented here examine diverse forms of reportable content. There are false voting instructions circulating on platforms, despite being a content type particularly suited for moderation (and removal) during the run-up to elections. There is algorithmic manipulation that seeks to groom recommendation systems, search engines and generative AI platforms to surface certain content when queried or prompted. There are cases of

coordinated behaviour through the rapid sharing of the same content by multiple, linked accounts to boost its visibility. We found a set of memetic, hateful conspiracy videos where content creators vary and add to them, trying to make their latest contributions viral. We also analysed such hateful content related to a key election issue across platforms with 'For You' algorithms, examining the extent to which each platform demotes it or allows it to be recommended.

The following studies are test cases about such reportable content, or posts by users that break platform rules, are reported or in some instances cannot be reported given the current platform mechanisms. The cases concern X/Twitter, TikTok, Instagram, YouTube and Telegram, platforms that have come under recent scrutiny during election campaigns in a series of European countries. One test case for X/Twitter concerns a false report that this year a citizen can cast two votes. For TikTok the case concerns a set of videos depicting Europe after 'Islamisation' or the 'great replacement' when the said white population has been displaced by a Muslim population. The presence or demotion of this content is also compared across four, highly personalised 'For You' algorithms (TikTok, X/Twitter, Instagram and YouTube). The Telegram case is about the automated amplification of Dutch political content by Russian sources, creating a false amplification 'infrastructure'. This is also a case of search engine and generative AI manipulation where a network of Russian sources shares political content that is expected to surface when searched for or prompted. We also report on an instance of coordinated inauthentic behaviour on the part of troll personas seeking to influence the Finnish parliamentary vote to join NATO.

For the X/Twitter and TikTok cases as well as the comparison of the For You algorithms, the content was reported, and where available the automated decisions taken by the platform about non-violation disputed. For the algorithmic manipulation, AI seeding as well as the example of coordinated inauthentic behaviour, these could not be reported in any straightforward manner, given the absence of a platform category for it to be flagged or a search engine and AI platform reporting system.

The absence of such reporting mechanisms for influence operations points up a selectivity in what the platforms allow to be reported through the interface, and the concomitant need to make these findings known in other manners, be it through national or EU-level reporting mechanisms, journalistic reporting or platform backchannels. How to let the platforms know?

Certain of our findings are counter-intuitive given an expectation that rule-offending content would be acted upon or could be reported in the first place. In the case of X/Twitter the hateful material remained online, as did most of the posts with false voting instructions, even though similar content had been removed previously.

The difference between the two reporting outcomes lies in their source; action previously had taken on a report by a trusted flagger (a national Ministry)

concerning the two-party voting instruction but not, in our case, from a regular user.

In the moderation auditing research on TikTok, the researcher reporting resulted in limited action by the platform, including a removal and a demotion. The moderation, however, was uneven. Whilst all of the same 'great replacement theory' genre, expressly in violation of platform rules, very few of these videos were moderated. The automated and manual systems were also occasionally in disagreement, indicating a misalignment between the automated as well as manual review systems.

For Telegram the researchers concluded that the content they found could not be reported, given not only Telegram's avoidance of the EU designation of a VLOP (a Very Large Online Platform according to the Digital Services Act), which would make it liable to stronger regulation and monitoring, but also because there is no category available for reporting FIMI. That is, for the infrastructural grooming of search engines and AI platforms grooming case, there is no reporting system in place for users.

The same could be said of the coordinated inauthentic network on X/Twitter seeking to influence a key parliamentary vote. One could seemingly report the users though as we discuss below there is no category corresponding to their violating behaviour and no open text field to describe the activity.

Each analysis includes implications. For all platforms there is the question of whether users have sufficient reporting authority. The question is posed not only because of the lack of action taken concerning the reportable content but also the inability to report certain user behaviours. For X/Twitter, researchers could not report a crucial content type, and there was no method available for appealing a decision. For TikTok the issue is the efficacy of automated content moderation generally as well as a misalignment between automated and manual review. For Telegram the findings point to not only the absence of reporting mechanisms but the presence of practices indicative of infrastructural, algorithmic manipulation. Given that they are perhaps more on an infrastructural level, search and generative AI manipulation and interference are not straightforward reporting types.

The reporting mechanisms available are oriented not only to content types but also to users. For the kinds of reportable behaviours found, however, the categories are not in place. Reporting a set of users as engaging in coordinated inauthentic behaviour using fake accounts, for example, appears unlikely or burdensome at best. One can report 'impersonation' but not a fake account. On TikTok, reporting impersonation has the secondary, sub-categorical request of impersonating 'me' or a 'celebrity' only. On X/Twitter the points of departure are slightly more expansive but still without the ability to report inauthenticity.

The question of which behaviours are reportable and how to perhaps work around a lack of specific category also raises the overall issue of the kind of labour involved in reporting. As said, when reporting content or a user one

chooses a rationale from a drop-down menu corresponding to platform rules and presses send. On certain platforms (such as TikTok), if one disputes a ruling there is space to appeal. It includes a free text field that allows to user to provide reasoning for the dispute, which presumably triggers a manual review. The reporting user can monitor progress within the interface through a list of numbered videos or users, but it is not designed for systematic reporting or research, given the level of detail and the formatting.

In the TikTok analysis of extreme mimetic videos the question of user-assisted content moderation includes an additional point about the algorithmic environment in which such assistance occurs. In 'For You' environments recommendation feeds are based on user preferences. Presumably, users would be interested in the content fed rather than be confronted with posts to be reported. The unlikelihood of seeing reportable content decreases the efficacy of user-assisted reporting.

It also might lead to considering other clean-up approaches such as search parties or edit-a-thons, where volunteers come together for a short period of time (in a data sprint, for example) to find and report rule-breaking content. How such dedicated, mass reporting would be treated by a platform remains an open question, especially if the users create new accounts for the occasion. Mass reporting may be viewed as a form of 'brigading', where a group of users join forces to harass rather than pitch in to help the platform moderate content.

References

Bradshaw, S., & Howard, P. N. (2018). The global organization of social media disinformation campaigns. *Journal of International Affairs*, 71(1.5), 23-32.

Chiarella, M. L. (2023). Digital markets act (DMA) and digital services act (DSA): New rules for the EU digital environment. *Athens JL*, 9, 33.

Hiltunen, M. (2021). Online political advertising and disinformation during elections: Regulatory framework in the eu and member states. *Helsinki Legal Studies Research Paper*, (68).

Malhotra, P., Zhong, R., Kuan, V., Panatula, G., Weng, M., Bras, A., ... & Zhang, A. (2023). User experiences and needs when responding to misinformation on social media. *Harvard Kennedy School Misinformation Review*.

Proto, L., Lamoso-González, P., & García, L. B. (2025). The EU's FIMI Turn: How the European Union External Action Service Reframed the Disinformation Fight. *Media and Communication*, 13.

Sekwenz, M. T., Gsenger, R., Stocker, V., Görnemann, E., Talypova, D., Parkin, S., ... & Smaragdakis, G. (2025, June). Can't LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work* (pp. 1-12).

Udupa, S. (2019). Digital disinformation and election integrity: Benchmarks for regulation. *Economic and Political Weekly*, 54(51), 1-8.

Vliegenthart, R., Van Ham, C., Kruikemeier, S., & Jacobs, K. (2024). A matter of misunderstanding? Explaining (mis) perceptions of electoral integrity across 25 different nations. *Public Opinion Quarterly*, 88(SI), 495-515.

The two-party vote: Reporting election misinformation on social media platforms

Stijn Peeters

Introduction

In the lead-up to the Dutch parliamentary elections on 28 October 2025, a meme circulated on social media regarding how one was to cast a vote for GroenLinks-PvdA – a new party emerging from a merger of GroenLinks and PvdA, two established left-wing parties. The meme claimed that to cast a vote for the party, one would need to fill two boxes on the ballot paper – one for both parties. In fact, this would make the vote invalid, as it is only counted if exactly one box is filled. As such, the meme could be seen as election-related misinformation, a category of content banned on many social media platforms and was designated as such by fact checkers. Platforms would then be expected to moderate such content if they become aware of it – but do they?

The meme (see Figure 1 for an example) is a version of an older 'meta-joke' regarding the voting process; another popular variation alleges that the elections have been spread over two days; one day for the side of the political spectrum the teller of the joke supports, and another one for the other side. The latter would then, of course, be at a later, non-election day. In many cases these jokes will be understood as such, or as satire, and those who spread it may often do so with this understanding in mind, rather than to mislead voters.

However, in this case the joke had a kernel of plausibility, given the fact that GroenLinks-PvdA truly is a merger of two parties that could previously be voted for separately. Combined with a broader concern about election-related misinformation, this prompted several 'fact checks' debunking the meme. According to reporting by AFP (Pauwels 2025), an international news agency, the Dutch Ministry of Internal Affairs also contacted social media platforms to ask them to flag such content due to its potentially misleading nature.

This followed earlier, similar requests from the Ministry during the previous election cycle in 2023 (Ministerie van BZK 2024), when a version of the same meme was circulating. In both cases, the Ministry used its status as a 'trusted flagger' – a designation introduced by the EU's Digital Services Act that requires platforms to treat reports with higher priority. Under the DSA, a trusted flagger can flag content but does not have the power to make moderation decisions; the decision of whether something is to be moderated always remains with the platform.

The AFP report mentions two specific posts that were flagged this way: one on Facebook, one on X. The Facebook post is no longer available (visiting it shows a generic “This content isn’t available right now” note). The X post is available but has been flagged by the platform with a notice saying, “Visibility limited: this post may violate X’s rules against Civic Integrity” (see Figure 1 for a similar post with the same label).

Figure 1. A post on X containing the ‘double vote’ meme, labeled as having its visibility limited and potentially violating X’s policies. Post text has been paraphrased. The text translates to “ADVICE FOR #GROENLINKSPVDA VOTERS... voters of #groenlinksPvdA may check TWO boxes on the ballot paper... @geertwilderspvv agrees with this !!”

As such, the meme can be considered ‘reportable’; it is an example of content that violates the platforms’ policy, which has been reported as such, and indeed flagged or moderated by the platform. Nevertheless, various other versions of the same meme – sometimes simply copy-and-pasted from the original – also circulated, and had not been reported to the platform, or in any case not been flagged. As these would be in violation of the same policies, the goal in this case study was to confirm whether reporting these posts via another reporting mechanism – the one available to ‘ordinary’ users – would likewise result in moderation.

The meme in question may seem relatively innocent, but election-related misinformation has been demonstrated to be a major issue facing social media platforms. This is a global phenomenon, and previous work has observed political and election-related misinformation to be prevalent on various social media in election campaigns in Brazil (Santini et al. 2021), Nigeria (Igwebuike &

Chimuanya 2020), Indonesia (Subekti et al. 2025), the United States of America (Dehlinger & Scala 2024), and the European Union (Casero-Ripolles et al. 2025), among others.

The issue is as widespread as it is varied; misinformation comes in many forms, from memes such as the one this case study concerns to targeted campaigns by governmental actors, also known as *foreign information manipulation and interference* (FIMI; see e.g. Willaert & Tuters 2025). Whether a given example of misinformation is the one or the other is not always easy to tell; targeted content or misinformation campaigns 'resonate' on platforms and can be spread by people unaware of its original source and origins, which may remain obscure (Hagen et al. 2025). In any case, that election-related misinformation is an issue is clear to both social media platforms and outside observers, and the meme at hand provides an opportunity to study how platforms handle its moderation.

Platform policies

To do so, we looked at three platforms: X, Instagram, and Facebook. X and Facebook have been demonstrated, as discussed above, to consider the meme in question worthy of moderation when flagged. Instagram, as a platform with the same owner as Facebook (i.e., Meta) may be expected to have a similar content moderation policy.

For all three platforms, we collected 10 posts published in October 2025 containing (a variation of) the 'vote twice' meme. For X, we collected a second set of posts from 2023, as previous reporting indicates that the meme was considered reportable back then as well, and we were interested in whether X would moderate these posts 'after the fact'.

All three platforms explicitly disallow election-related misinformation:

- **Facebook** claims on its 'Our approach to elections' page (Meta 2023) that "[w]e remove content that attempts to interfere with voting, such as incorrect voting information". Additionally, on its page describing 'community standards' (Meta n.d.) concerning Misinformation, it notes that "we remove misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in those processes" including "misinformation about the dates, locations, times, and methods for voting, voter registration, or census participation".
- **Instagram** has mostly adopted the same policies as Facebook; links to its 'Community guidelines' point towards the same documents as those that apply to Facebook, which are collected in the 'Meta Transparency Center'. As such, Instagram is considered to follow the same policies as Facebook.
- **X**, in a section in its policy documents titled 'Civic integrity' (X 2025), notes that it does not allow "advancing verifiably false or misleading information about how to participate in an election or other civic process" including "misleading claims that cause confusion about the established

> laws, regulations, procedures, and methods of a civic process", where political elections are explicitly noted as an example of such a civic process. This policy document is also linked to in the warning label added to the post that was mentioned by the AFP as having been reported by the Ministry, suggesting that it is indeed the relevant policy here.

Taken at face value this suggests that any election-related misinformation, including the 'two-party vote' meme, would be removed from the platforms. Content moderation on X has however been reported to be generally "unclear and inconsistent" (De Keulenaar 2025). This is especially true in the wake of its acquisition in 2021 by tech billionaire Elon Musk, who has adopted a particularly belligerent and litigious approach to legislative efforts to promote content moderation (see e.g. Dang 2023; De Guzman 2025); Musk has called content moderation "a propaganda word for censorship" (De Guzman 2025).

Meta, meanwhile, is perhaps less openly hostile towards content moderation, but nevertheless has a complicated relationship with the practice. While it works with trusted third-party fact checkers and has set up a nominally independent 'Oversight Board' to adjudicate moderation disputes, in 2025 it ended its cooperation with fact-checkers in the USA and announced a shift to a 'community notes'-oriented system of moderation (Kaplan 2025) calling its current content moderation "censorship", a move that was then "sharply rebuked" by the Meta-funded Oversight Board for its potential negative human rights impact (Paul & Wang 2025). EU laws require Meta to keep working with third-party fact checkers as well as other "governance entities" in that region (Leerssen 2024, p.18) – as evident in the case study at hand. But these recent moves suggest that the company may be moving rather towards a stance on content moderation not too dissimilar from Musk's and X's. In any case these relatively recent policy shifts suggest that an empirical study of the platforms' responses to user-reported banned content is timely.

Method

The goal was to report posts that re-used the 'two-party vote' meme that had been flagged according to the AFP report verbatim or used a close variation of it. This included the use of a photo of a ballot paper (see Figure 1). To find such posts, a mix of query strategies was used. The platform's own search feature could be used to find posts containing relevant phrases such as "twee kandidaten" ("*two candidates"*) or "twee bolletjes" (*"two bullet points"*). On Instagram, which has no platform-wide text search feature, Google was used with a site search query, e.g. [twee bolletjes site:Instagram.com].

While election-related misinformation is defined quite specifically as not allowed on all three platforms, they do not actually offer a way to report content as violating this specific policy. All three platforms' reporting mechanisms – in any case those available to an ordinary user – only allow reporting within broader categories such as 'Privacy' or 'Child Safety'. The most relevant reporting pathways at the time of the analysis were as follows:

- **Facebook**: Report Post → Scam, fraud, or false information → Shares false information
- **Instagram**: Report → False information
- **X:** Report Post → Spam

When used in the EU, X offers an additional reporting link, titled 'Report EU illegal content'. In contrast with the 'ordinary' reporting option, following this link takes the reporter to a form in which more specific information can be provided to motivate the report. Here we chose to report the posts for 'Negative effects on civic discourse or elections', after which a more specific subcategory can be chosen. Two of these were equally relevant, and so five posts were reported for 'Violation of EU law relevant to civic discourse or elections' and five for 'Misinformation, disinformation, foreign information manipulation and interference'.

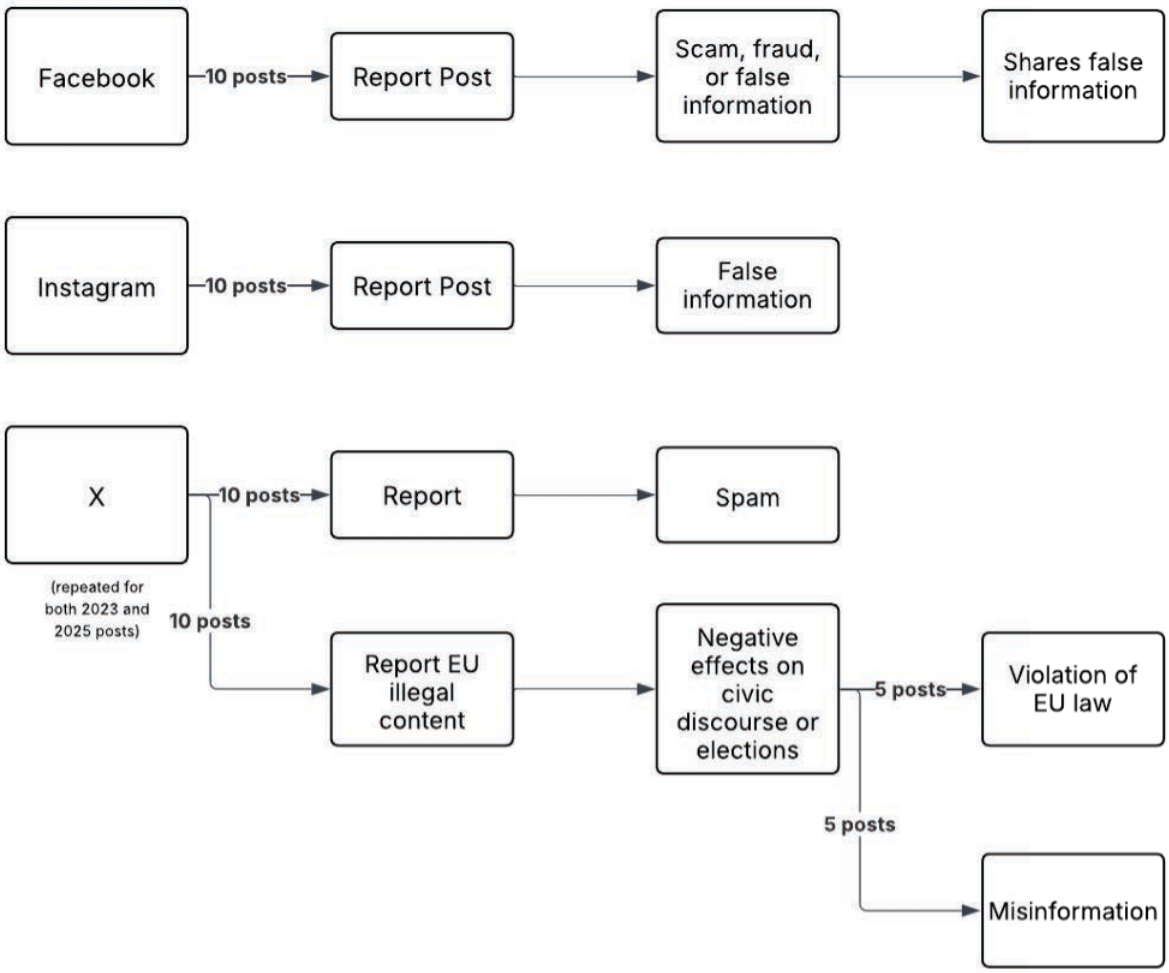


Figure 2. Reporting mechanisms used.

The reporting pathways used are visualized in Figure 2. The goal was to test each of these with a set of 10 posts containing the 'double vote' meme and then observe 1) whether any response to the report would be sent by the platform and

2) whether the reported content would be removed, labelled, or otherwise moderated. Since our goal is not to assess the prevalence of the meme, but rather the platforms' moderation practices and responses to user-reported banned content, the sample size can be relatively low. All posts are unambiguous examples of the meme, and by the platforms' own policies as well as precedents of their application should be moderated. A small sample is thus sufficient to observe the moderation practices in this context.

All posts were reported on 23 October 2025, one day before the elections. We then monitored the reported posts as well as any responses to our reports from the platforms for a period of four weeks, until 25 November 2025.

Results

In summary, reporting the meme had no clear effects on the availability of the posts. Posts were checked one week after reporting and then again four weeks after reporting.

- On **Instagram**, all posts were still online as of 25 November 2025, with no warning labels or other indications that action was taken. While Instagram's help pages claim that the status of reported posts can be followed in the website's Settings page (Instagram n.d.), no records of the reports could be found here.
- On **Facebook**, some posts had been labelled or disappeared one week after reporting (Table 1). However, those that had been labelled noted that they had been flagged by dpa-Faktencheck, a fact-checking outfit part of the Deutsche Presse-Agentur that works with Meta directly; it is thus unlikely that the posts had been labelled following our reporting. We received no responses to or notices regarding our reports, and reports were not visible in the site's settings though documentation says this should be the case (Facebook n.d.). Some posts had become unavailable without a stated reason. It is possible this happened in response to the reporting, though other, identical, reported posts were still available, and the other posts that were removed for policy violations (following a third-party fact-check) were clearly labelled as such.
- On **X**, a single post from October 2025 had been labelled as potentially violating X's rules (see Figure 1). Two other posts were no longer available as its author's account had been suspended from the platform (i.e. all their posts had disappeared, not just the one reported). It is unclear if the labelling of posts or account suspensions (see Table 2) were a result of our reporting. That many identical or near-identical posts that we also reported in the same way are still available suggests that this is not the case. For 11 out of 20 posts reported with the 'EU illegal content' feature, we received a rejection of the report within an hour after reporting. For the other reports, we received no update or response.

Status as of 25 Nov 2025	n	Notes
Available	5	
Available, but hidden with a flagging notice	2	Flagged by 'dpa-Faktencheck', a trusted third-party flagger
Unavailable	3	Displays 'This content isn't available right now'. No indication whether it was removed, hidden by the owner, or moderated.

Table 1. Status of reported items on Facebook as of 25 November 2025 (four weeks after reporting).

Status as of 25 Nov 2025	n	Notes
Posted in 2025, reported via standard reporting features		
Available	10	
Posted in 2025, reported via the 'Report EU illegal content' feature as misinformation		
Available	4	
Labelled as violating rules	1	Post had 22 views and 1 like.
Posted in 2025, reported via the 'Report EU illegal content' feature as breaking EU laws		
Available	4	
Author account suspended	1	
Posted in 2023, reported via standard reporting features		
Available	9	
Author account suspended	1	
Posted in 2023, reported via the 'Report EU illegal content' feature as misinformation		
Available	5	
Posted in 2023, reported via the 'Report EU illegal content' feature as breaking EU laws		
Available	4	
Author account suspended	1	

Table 2. Status of reported items on X as of 25 November 2025 (four weeks after reporting).

Discussion

The results suggest that reporting content, even if it has been demonstrated to violate platform rules, has no clear effect if the reporting is done via the user reporting mechanism. There are obvious limitations to this analysis' approach

that require some reservations regarding this claim. The reports concerned variations on a single post; platforms may more readily act on reports on other types of content, such as CSAM or hate speech.

Nevertheless, all platforms included have clearly banned election misinformation and express a concern for election integrity in their policy documents. That content which violates this policy – judging by the platforms' own actions – is then not moderated after reporting it the day before the election this misinformation concerns, is cause for concern. While a longitudinal analysis would be necessary to know if this is a new phenomenon, X's and Meta's recent shift towards a 'light touch' approach to moderation suggests that the lack of response to user reports may, at least currently, be by design.

The sample size of this analysis – 10 posts per reporting method – was relatively low, but since all 10 posts are clear and close variations on a demonstrably 'reportable' post, it is unlikely that a higher sample size would have led to different results. If the platforms would consistently follow their own policies regardless of reporting method, one would expect these posts to have been moderated, regardless of how many were reported.

That this did not happen indicated that either these reports are reviewed but not considered actionable; or that the reports were automatically rejected or ignored. That similar posts were labelled because of flagging from other sources (the Dutch Ministry of Internal Affairs), and dpa-Faktencheck) indicates that this content is indeed reportable – but clearly, it matters who is doing the reporting.

Conclusion

We find that on various social media, reporting content that has been proven to be 'reportable' - that is, other instances of the same content have been flagged or removed by the platform - are not flagged or removed after reporting. This suggests that while such content can be and is occasionally moderated by the platform, such moderation does not happen (primarily) because of 'crowd-sourced' user reports, but rather via other mechanisms such as third-party fact checkers, 'trusted flaggers', or algorithmic moderation. Meta's abandonment of this model in the US context – but not in the EU where it is required by law – suggests that in the European Union, this baseline of moderation is in place more due to the legal context than the platforms' own concern for the spread of election-related misinformation.

Having reported proven reportable content through a variety of reporting mechanisms on three different social media platforms, it seems likely that such user reporting is largely ineffective or in any case not a viable mechanism for flagging such content in a timely manner. The content that was reported here – misinformation concerning voting methods for the 2025 Dutch parliamentary elections – violated all these platform's policies and was reported the day before these elections. It is nevertheless largely still online four weeks later, and insofar as posts were labelled or taken offline, this seems to have been due to other factors such as reports from third-party fact checkers or the suspension of the post author's account for other reasons.

While the content discussed here – a humorous meme – is perhaps not a significant danger to the electoral process, this nevertheless gives cause for worry regarding the platforms' stated concern about safeguarding civic processes and limiting the spread of misinformation, election-related or otherwise. (Systemic) risks concerning the spread and weaponization of such misinformation have been demonstrated time and time again; that platforms seem reluctant to proactively address this through providing effective reporting mechanisms is thus problematic.

Post scriptum

As of November 2025, X includes a new category for which content can be reported: 'Civic Integrity', summarized as 'misleading content related to voter participation, suppression, or intimidation in elections and other civic processes'. This category, which would have perfectly fit this case study, was not available at the time.

Acknowledgements

Thanks to Abdoul Yerbanga and Rong Sun for their help with the collecting and reporting of relevant posts.

References

Casero-Ripollés, Andreu, Laura Alonso-Muñoz, and Diana Moret-Soler. 2025. "Spreading False Content in Political Campaigns: Disinformation in the 2024 European Parliament Elections." *Media and Communication* 13 (0). https://doi.org/10.17645/mac.9525.

Dang, Sheila. 2023. "Musk-Owned X's Content Moderation Shift Complicated Effort to Win Back Brands." Technology. *Reuters*, September 7. https://www.reuters.com/technology/musk-owned-xs-content-moderation-shift-complicated-effort-win-back-brands-former-2023-09-07/.

Delinger, Josh, and Nathalie M. Scala. 2024. "From Misinformation to Trust: Safeguarding the 2024 U.S. Presidential Election." *ORMS Today* 51 (3). https://pubsonline.informs.org/do/10.1287/orms.2024.03.03/full/.

Facebook. n.d. "Can I Check the Status of Something I've Reported to Facebook or Cancel a Report?" Facebook Help Center. Accessed November 25, 2025. https://www.facebook.com/help/338745752851127/.

Guzman, Chad de. 2025. "Musk's X Sues N.Y. in Latest Battle Over Content Moderation." TIME, June 18. https://time.com/7295402/elon-musk-x-new-york-lawsuit-free-speech-content-moderation/.

Hagen, Sal, Daniël de Zeeuw, and Tommaso Venturini. 2025. "Digital Rhythmanalysis: Studying Memetic and Affective Rhythms on the Post-Viral Web." *Platforms & Society* 2 (December): 29768624251394967. https://doi.org/10.1177/29768624251394967.

Igwebuike, Ebuka Elias, and Lily Chimuanya. 2021. “Legitimating Falsehood in Social Media: A Discourse Analysis of Political Fake News.” *Discourse & Communication* 15 (1): 42–58. https://doi.org/10.1177/1750481320961659.

Instagram. n.d. “Check the Status of Something You’ve Reported to Instagram.” Instagram Help Center. Accessed November 25, 2025. https://www.facebook.com/help/instagram/484700649522086/.

Kaplan, Joel. 2025. “More Speech and Fewer Mistakes.” *Meta Newsroom*, January 7. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

Keulenaar, Emillie de. 2025. “From Twitter to X: Demotion, Community Notes and the Apparent Shift from Adjudication to Consensus-Building.” SSRN Scholarly Paper No. 5165083. Social Science Research Network, March 1. https://doi.org/10.2139/ssrn.5165083.

Leerssen, Paddy. 2024. “Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act.” *Weizenbaum Journal of the Digital Society* 4 (2). https://doi.org/10.34669/wi.wjds/4.2.3.

Meta. 2023. “Our Approach to Elections.” Meta Transparency Center, November 27. https://transparency.meta.com/features/approach-to-elections/.

Meta. 2025. “How Fact-Checking Works.” Meta Transparency Center, April 7. https://transparency.meta.com/features/how-fact-checking-works/.

Meta. n.d. “Community Standards.” Meta Transparency Center. Accessed February 23, 2026. https://transparency.meta.com/policies/community-standards/.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. 2024. *Inzet Trusted Flagger Status BZK.* Rapport. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. https://www.rijksoverheid.nl/documenten/rapporten/2024/03/22/bijlage-10-rapport-inzet-trusted-flagger-status-bzk-tk23.

Paul, Katie, Echo Wang, and Echo Wang. 2025. “Meta’s Oversight Board Rebukes Company over Policy Overhaul.” Boards, Policy & Regulation. *Reuters*, April 23. https://www.reuters.com/sustainability/boards-policy-regulation/metas-oversight-board-rebukes-company-over-policy-overhaul-2025-04-23/.

Pauwels, Lisa. 2025. “Kiezers van GroenLinks-PvdA moeten géén twee vakjes kleuren, dat maakt de stem ongeldig | Factcheck Nederland.” AFP Factcheck Nederland, AFP, October 24. https://factchecknederland.afp.com/doc.afp.com.79MY6NQ.

Santini, Rose Marie, Giulia Tucci, Débora Salles, and Alda Rosana D. de Almeida. 2021. “Do You Believe in Fake After All?” In *Politics of Disinformation*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119743347.ch4.

Subekti, Dimas, M. Yusuf, Maratun Saadah, and Makmun Wahid. 2025. “Social Media and Disinformation for Candidates: The Evidence in the 2024 Indonesian Presidential Election.” *Frontiers in Political Science* 7 (July). https://doi.org/10.3389/fpos.2025.1625535.

Willaert, Tom, and Marc Tuters. 2025. “From Denazification to the Golden Billion: An Inductive Analysis of the Kremlin’s Weaponisation of Digital Diplomacy on Telegram.” *Humanities and Social Sciences Communications* 12 (1): 989. https://doi.org/10.1057/s41599-025-05382-x.

X. 2025. “X’s Civic Integrity Policy.” X Help. https://help.x.com/en/rules-and-policies/election-integrity-policy.

Extreme memetic videos on TikTok and a test of their moderation

Richard Rogers

The piece reports upon a test case of content moderation by the social media platform, TikTok, of hateful conspiracy videos targeting Muslims. The videos in question depict a European future when in 2050 cities have been 'Islamised'. The paper examines how creators position these videos as content-as-usual, with platform hashtag and descriptor norms referencing memetic user culture and virality efforts. Subsequently, in the test case, the videos are reported to the platform through the content moderation reporting function as violating its hateful conspiracy policy and after (mostly) non-violation decisions were made the decisions are appealed. In all, it was found that there is an unevenness in moderation decisions between automated and manual review. The implications of these findings are discussed in terms of automated slippage and human-machine misalignment as well as the user reporting burden assumed in an algorithmic environment promoting personalised content. It concludes with a call for moderation auditing to test platform claims about the efficacy of automated moderation systems.

Introduction: Islamisation as narrative theme and TikTok video meme

This study is part of a larger project about the moderation of (foreign) misinformation and other election-related content that impinges upon social media platform rules. Conducted in the run-up to national elections in the Netherlands, it begins with the identification of a 'disinformation narrative' (related to the election issue of migration) concerning the Islamisation of Europe, its debunking, its origins as well as its continuing online life on TikTok, the popular social media platform for short videos. It discusses how a series of Islamisation videos set in European cities in the year 2050 – Amsterdam, 2050, Barcelona, 2050, Milan, 2050 and so forth – portray devastation after the religious and cultural shift. Through hashtag usage and video descriptions, they could be considered mimetically embedded in TikTok's user culture; they also remain online, despite ostensibly breaking platform rules concerning hateful conspiracy narratives. Subsequently, these videos are reported to the platform to conduct an audit of the platform's content moderation systems, which lead to findings about an unevenness in content moderation, a misalignment between automated and manual review and the need for user assistance in platform moderation tasks.

The link between disinformation and hateful conspiracy arising in Europe 2050 was raised some years ago by EUvsDisinfo, the non-governmental organisation, when it posted a fact check story about a trend recorded in online news stories. It concerned the alleged coming of 'Islamisation', or how a rising Muslim population in Europe would supplant Europeans and their values by 2050,

including their Christian orientation, extrapolating from statistics published by the Pew Research Center (EUvsDisinfo, 2019). The story they fact checked, entitled ‘With a crescent moon on the Christmas tree’, was published in the same year that departing Syrian war refugees peaked; it also was of Russian origin, lending it an additional dimension. In debunking claims of so-called population replacement, or a majority Muslim population supplanting European Christians by the year 2050, EUvsDisinfo also wrote that it was a ‘pro-Kremlin’ narrative. It relates not only to a threat of Islamisation but also to the preservation of older moral values.

There are stories with related themes in EUvsDisinfo’s larger case book on Islamisation which date further back, providing a fuller picture. The EU’s allowance of the influx of non-Europeans relates to a complex of alleged European problems, revolving around the weakening or ‘emasculation’ of Europe and its peoples, be it through the intermingling of religions, the rise of LGBTQ+ rights as well as policies preaching ‘tolerance’.

In each instance, these stories, especially when from Russian sources, directly or indirectly, are deemed ‘disinformation’ efforts, which EUvsDisinfo describes as a revitalised practice, originating in the 20th century as ‘infection operations’ that seed and cultivate discord for geopolitical purposes (2017). While it assumes many forms, here the elements of the narrative together are considered conspiracy theory, where the EU or European elites more broadly are the influential agents operating in the dismantling of European (and Western) values through allowing in migrants and refugees that emasculate and Islamise.

The story of a coming demographic crisis is rooted in a much longer history with several branches, though the Islamisation premise is exemplified by the contemporary thought of Eurabicists who argue that Europeans, in an Islamised state, would suffer the consequences of dhimmitude, or a system of submission to Muslims by non-Muslims living in an Islamic state. Referring to it as the submission discourse, scholars have described it as an “international theoretical site for Islamophobic intellectuals” (Kattar, 2025).

A more recent variant is the great replacement theory, put forward initially in response to the presence of Muslims in France, where so-called institutional elites are described as ‘replacists’, substituting Europeans with non-Europeans, thereby posing a threat to the ethnonational populations through what is referred to as ‘white genocide’. Characterised by researchers as an extremist, anti-immigration and conspiratorial narrative (Dennison & Kustov, 2025), it is a story that has reached the highest levels of the US government, where President Trump has called for its cessation in South Africa and put forward immigration policy plans that would favour white people (Kanno-Youngs & Aleaziz, 2025).

Here is where the disinformation efforts intersect with nationalist and nativist politics, contributing to a broader geopolitical drive to further cultural divisions in Europe as well as the West. They enable political points of view that exhibit less tolerance towards the outsider and the other and more sympathy with

morally conservative values more broadly. It fits with what researchers term the ‘threatened values’ narrative (Enerud, 2022).

This larger narrative tide is carried and furthered by internet influencers and media personalities, often taking up the disinformation talking points; there are also contributions from their audiences and other adherents and users, whether in it for ideological purposes, in-group amusement, engagement farming and/or content monetisation.

They participate in its cultural production which on TikTok are short videos that range, in this case, from depictions of European cities in 2050 after Islamisation or maps of European regions gradually shaded or absorbed to portray the religious and cultural takeover. Scholars have dubbed these bottom-up contributions to the disinformation and conspiracy narratives as ‘participatory propaganda’ when intentional and ‘ambient propaganda’ when unwitting (Chernobrov, 2025; Tuters & Noorderbos, 2023).

The research presented here concerns the everyday presence of these narratives on TikTok, particularly through the Europe 2050 videos that were collected just prior to the 2025 Dutch parliamentary elections, where migration has been a key issue, given the attention given to it by a leading party, the Freedom Party, and its leader Geert Wilders, who had topped the polls in the previous election. During the campaign period, Wilders, for his part, also posted a Netherlands 2050 video (on X/Twitter) with the now-and-then storyline of the coming consequences of Islamisation including trash-strewn streets, bazaars overtaking markets and mosques lining the skyline.

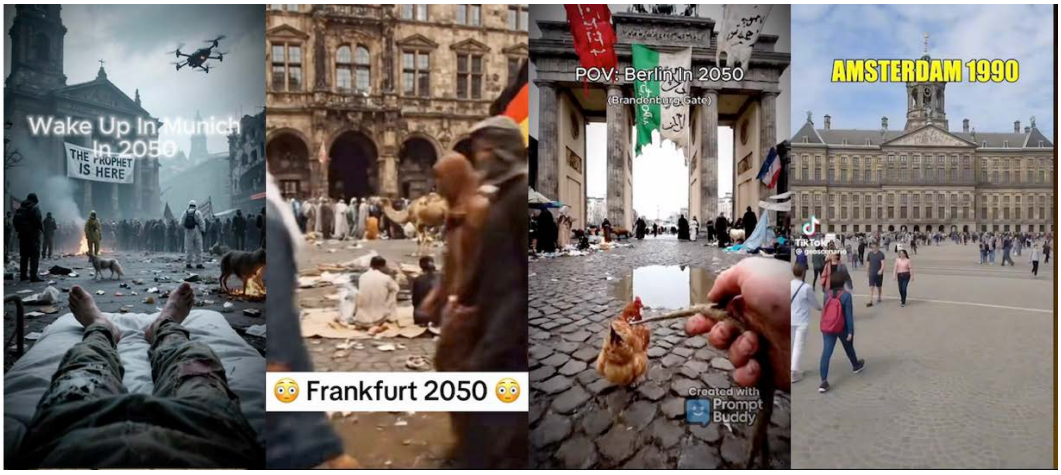


Figure 1: Select thumbnails of Europe 2050 TikTok videos. November 2025.

Some of the Europe 2050 videos are AI-generated or AI-enhanced, often taking the first-person point of view, allowing the embedded viewer to experience feeling foreign in one’s own land through immersion (see Figure 1). These are either walkthroughs through a barely recognisable city or moments after waking up some decades later. Others, with the maps, have a more factual, documentary style, making an alternative factual argument – rather than conjuring the experience – of being culturally replaced.

From the literature these videos could be said to be created by 'imitation publics' participating in meme culture, where one contributes to a TikTok meme through using a sound that is trending or an identifiable style that is circulating (Zulli & Zulli, 2020). By meme is meant here a collection of videos that imitate content, form and stance, varying them as the 'additive content' multiples and spreads (Shiffman, 2013). The videos are also part of the 'replacist' online cultural space, and its dark participatory culture (Quandt, 2018). Indeed, they contribute to the memetic trend through acknowledging that in their descriptions and hashtags; a small set is more outwardly political in their orientation, explicitly referring to hard right European political parties.

Given how these videos are well embedded in TikTok culture, and how each new city depicted contributes to the larger theme of an Islamised Europe in 2050, ultimately, I make a series of points about how to think about TikTok's content moderation. First, it may be argued, that these videos are very much a part of everyday TikTok culture, where users strive to add to memes and make them more visible or trending in the For You algorithmic environment. At the same time, the videos display not just an ease with TikTok culture but a particular extreme politics vilifying and sullying Muslims. Their style could be dubbed extreme memetic politics, where video after video performs and provides access to an ethnonationalist point of view.

'Reportable content' and its moderation by TikTok

The substance of the videos is considered 'reportable content', or material that breaches platform policies on hate and is eligible for moderation. In the literature hate is often discussed as devoid of an agreed to definition for the purposes of content moderation and thereby debatable (Singhal et al., 2023). In practice, however, its initial determination has been given over to what is referred to as algorithmic content moderation (Gorwa et al., 2020), implemented in automated systems, deploying a suite of metrics that score content for a variety of measures, including hate. TikTok does not provide much detail on the mechanics of its detection, apart from writing that it uses automated systems to remove posts or reduce their visibility either for a younger age group or through making it ineligible for its For You Page (TikTok, 2025), which scholars have described as "recommending content aligned with users' interests and identities" and as its sticky recommendation system driving compulsive use (Cullen et al., 2025; Bhandari & Bimo, 2022). Human moderation takes place when there is some level of ambiguity from the automated scoring or when users appeal.

While the algorithmic moderation techniques employed are not well known, the categories of content moderated are described. At the time of writing, among the top-level categories of content to be moderated is 'safety and civility'. Falling under it is 'hate speech and hateful behaviour' towards 'protected groups' legally safeguarded from discrimination owing to attributes including race, religion, gender, or sexual orientation (TikTok, 2025). Under 'hate speech and hateful behaviour', TikTok provides specific examples germane to this research, particularly concerning content that contains 'hateful ideologies' including

‘white supremacy’ and ‘Islamophobia’ as well as ‘hateful conspiracies’ directed towards a protected group, including the ‘great replacement theory’ (TikTok, 2025).

The way TikTok’s community guidelines are put into practice has been the source of scholarly work, especially with respect to resultant content moderation through ‘visibility moderation’ or demotion (Zeng & Kaye, 2022). Some of the work came on the heels of leaked internal documents, revealing demotion practices for political content as well as for videos of ‘unattractive’ people and those with disabilities (Chen, 2019; Allyn et al., 2024). There are also studies on missing moderation as well as its circumvention through algospeak and other practices (Lookingbill & Le, 2024). User ‘folk theories’ of how the For You Page works also have been a source of scholarly attention (Klug et al., 2021; Felaco, 2025), including user frustration with being button-holed by the recommendation system (Vera & Ghosh, 2025). There is also work on moderation injustice (Are, 2025).

There is journalistic reporting about the labour involved in manual TikTok content moderation. These moderator experiences echo other investigations on what is referred to as commercial content moderation (Roberts, 2016; Gillespie, 2018). Moderation labour is difficult to study in practice, for non-disclosure contracts prevent content moderators from speaking about their work. In one anonymous recounting, it is described as mentally “soul-crushing” work, requiring rapid, complex decision-making about questionable content whose continual performance review makes workers prone to burn-out (Farah, 2023). Such findings point to the challenges of adjudication by individuals working as content moderators, who see the video content when flagged by automated systems or after an automated assessment is disputed, as we did.

Europe 2050 videos: Analytical starting points

To undertake the analysis of the moderation of (foreign) misinformation or other election-related content that violates platform rules, a group of researchers demarcated a source set of TikTok videos related to an issue animating political discourse in the Netherlands: migration. We first created keyword lists containing generic terms such as asylum seekers (asielzoekers in Dutch) as well as more colourful ones such as population replacement (omvolking) and we want our country back (wijeisenonslandterug), a slogan associated with a recent political rally in The Hague. We queried these keywords in the platform, and through co-occurrence analysis of the keywords and/or hashtags found in the descriptions of the videos, we added terms we had missed. Through new, enhanced queries, we collected some 1,000 video posts, which surfaced a select number of Europe 2050 videos about Islamisation (both first-person POV as well as the aforementioned documentary-style). We subsequently curated a collection of the Europe 2050 videos from the entire set, some twenty in all.

To approach the question of their fit in TikTok culture, we analysed how they are presented to other TikTok users in their descriptions and their hashtags. To do so, we created a co-hashtag graph, where hashtags related to the videos cluster thematically (see Figure 2).

While they may be embedded in TikTok's memetic user culture, their substance seems to clearly violate platform rules. Thus, we asked, does TikTok moderate these videos if reported? We reported the content to the platform, recorded its reporting and monitored takedowns or other moderation effects. After our reporting results in a notice by TikTok of non-violation, we dispute the company's assessment and add a note in the open text field describing why the video should be moderated. It reads:

> This video promotes the "Great Replacement" conspiracy theory, which falsely claims that certain racial or ethnic groups are being intentionally replaced. It spreads hate and fear toward immigrants and minority communities and encourages discrimination and hostility.

We then record the platform's response to the appeal (see Figure 3).

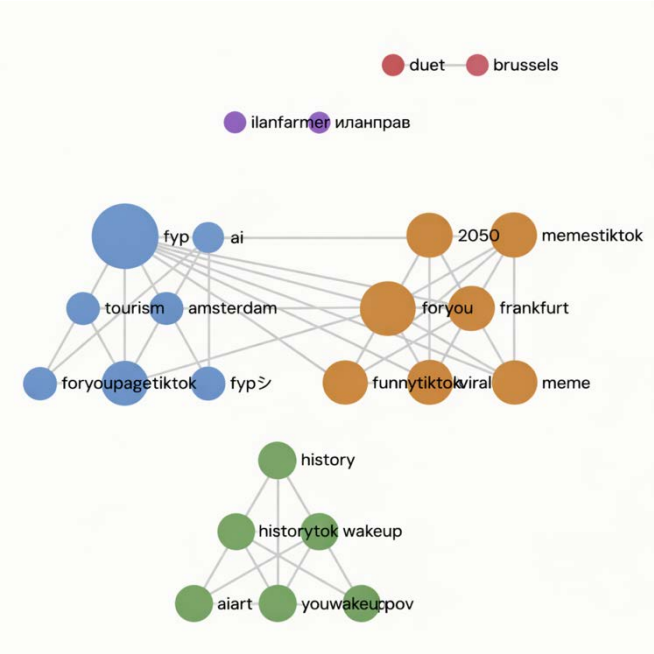


Figure 2: Graph of co-hashtag analysis of Europe 2050 videos. October 2025.

In a later step, after finding that TikTok removed the most popular video in our set after we appealed its initial non-violation finding, we asked why the video had not been caught by the automated moderation system in the first place. For some insight, we ran the transcript of the video through a content moderation scoring system that measures language inputted in it for 'toxicity', 'threat', 'profanity', 'identity attack' and other metrics. The purpose here is to gain a sense of whether it could be said to meet certain thresholds that would trigger moderation, were these metrics or similar ones used.

To sum up the approach of the work, we took to TikTok to study the key election-related issue of migration, ultimately making a collection of particularly poignant videos concerning Europe 2050 after so-called Islamisation which appears to be in violation of its community standards guidelines concerning hateful conspiracy. We reported them as violations of the policy, recorded their

reporting and studied how they were assessed by TikTok. We also ran a popular Europe 2050 video through a suite of metrics to assess its potential capacity to trigger moderation automatically.

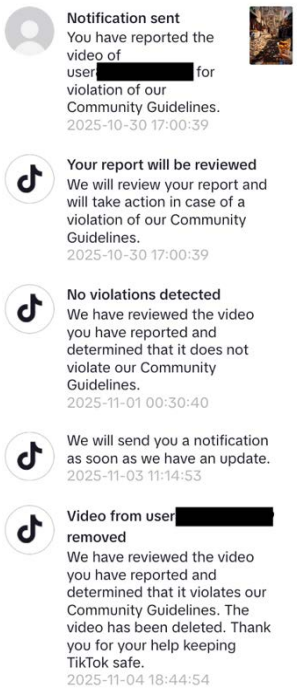


Figure 3: Screenshot of TikTok reporting log. 4 November 2025.

Findings: Extreme, memetic and politically charged For You videos

The findings concern the embedding of these Europe 2050 videos in TikTok memetic culture, their narrative style and their lack of moderation, despite meeting the definition of rule-breaking, the concomitant removal of other videos of the same genre and (in the case tested) scoring high for certain moderation metrics. We also found that TikTok removed a video after receiving appeals about their initial decisions, meaning there is occasional disagreement between the automated detection system and the human moderators.

To give some context about the popularity of the videos and the orientation of their creators, few in the set could be said to have reached a substantial level of reach and engagement; most have low view and like counts at the time of viewing. Some have been online for six months, the majority a month or two. The more popular ones could be classed as political advocacy; one calls for support of the Freedom Party in the Netherlands and others for the Vox Party in Spain and Alternative for Germany, respectively. The majority, however, is by content creators contributing to the meme and/or concerned with the same issue, largely without great followings.

With respect to how the videos are embedded in platform culture, the co-hashtag analysis we undertook shows the familiarity of the content creators with TikTok user practices through hashtag deployment; they add meme and virality hashtags rather than say a set of extreme terms such as we used at the outset such as replacement or Islamisation (see Figure 2). Examining the co-hashtag

graph, one set of related hashtags contains #fyp and similar ones related to TikTok's For You Page which the users hope to penetrate. Another set has to do with TikTok's memetic culture; creators tag videos with #meme, #funnyTikTokviral and similar. The third set has to do with a specific space or subculture on TikTok towards which the videos are oriented, called #historytok, with videos about historical subject matters, including in the style of the now-and-then as well as waking up in the future, with the hashtag, #youwakeup.

A brief discussion of the narrative and style of the videos themselves is in order. The Europe 2050 videos, as said, are largely of two types: the 'wake up' videos when one arouses from a slumber and looks out the window and scans the city, and first-person point-of view (POV) videos where one walks through the familiar squares and monuments recast as Islamised.

Each video depicts the future as (mainly) cities polluted, ill maintained, smoky, and populated by foreigners, mainly men wearing long, white garments and skull caps. There are puddles underfoot and open fires. Trash bobs in rivers, canals and fountains. Sheep and goats linger in squares and in commons. Markets have become bazaars, and mosques have supplanted churches. Signs are in Arabic. There is reference to Sharia law. Occasionally there are rickshaws and camels as forms of transportation as well as small river barges. There are no trains, cars, buses or other more modern forms of transport in sight. These are depictions of the future as regression and are acts of admonition, warning what is ahead unless action is taken. As a case in point, here is the transcript of Barcelona 2050, which TikTok left online after we reported it and after appeal. It was the most popular in our set.

> By 2050, Barcelona will belong to Morocco. I'm not saying this; the numbers prove it. Mass immigration and a lack of integration have led to an unprecedented demographic shift.
>
> Catalonia will no longer be Catalan or Castilian, but a territory with a completely new identity. The streets, schools, and culture will be dominated by Moroccan and Pakistani influence. Long-standing traditions will disappear, while Islam becomes the majority religion.
>
> And all this while politicians continue to ignore reality. In 30 years, Catalonia will be the most Islamised region in Western Europe. This isn't just a theory; it's the result of unchecked policies where the future of Catalans has been decided without their input.
>
> Is there still time to change it? Or is Catalonia's fate already written?

The video will have passed through the automated content moderation system. To gain a sense of whether the transcript would trigger typical, automated content moderation thresholds, we ran it through a moderation scoring test. We asked an LLM to score it on typical moderation metrics. It provided scores for "toxicity, insult, threat, profanity, identity attack, and inflammatory / fear mongering" (OpenAI, 2025; see Appendix). Of the metrics, it found the highest

score for identity attack (0.85-0.95 out of 1.0), which is a category of particular interest, given policies against hateful conspiracies towards 'protected groups'. The scoring also highlighted how the text was strongly directed towards religious and ethnic groups.

We also reported all videos through the standard TikTok reporting panel as in violation of the policy on hate. One of the twenty videos was removed, one was excised from the For You Page, and the others were deemed in non-violation according to the automated system reply. When these determinations were appealed, and an additional rationale added, the manual check reversed one decision. The other registered disputes resulted in TikTok upholding the original decision to keep the videos online (without any demotion).

Implications: Automated slippage, human-machine misalignment and the user reporting burden

In the following I discuss certain implications of the findings, particularly how these hateful conspiracy videos evade routine automated detection (when at least in one case moderation metrics indicate a violation). Another implication is the burden of moderation that is placed on TikTok users and the expectation that it may be met given the algorithmic environment.

This user reporting burden has implications for which videos ultimately will be moderated, since the algorithmic system – the For You Page – serves content to those most interested in it. That is, users served Europe 2050 videos presumably would be less likely to report the content, given their determined interest in them. Here the algorithmic recommendation system works against user-assisted moderation, making the overall moderation system less productive, particularly if the automated filtering does not trigger moderation in the first place.

In their policy concerning hate and harassment, TikTok does not allow videos that are "hateful conspiracies targeting a protected group, such as the "Great Replacement Theory" (TikTok, 2025). Here one could discuss the question about whether there is an exact or fuzzy match between the policy and the offending videos in a definitional sense, though certain of the videos in the same genre have been moderated automatically and manually (after reporting them) while others have not.

We do not have access to TikTok's own automated system for moderation, but moderation scoring would give a sense of how the language in the transcripts would fare when measured for such metrics as toxicity, identity attack, threat, and inflammatory/fearmongering. In the event, the metric 'identify attack' (of Muslims) scored high, calling into question how the videos slipped through, and indeed why they would not have triggered moderation.

When not caught by the automated moderation system, who will report these videos? Content reporters could well be scarce in the sense that a well-functioning algorithmic environment would serve the videos to users most interested in them, who in turn would be less likely to report them. Users who

would suggest that they be moderated would have to encounter them one way or another, perhaps more likely through search. In all, it could be argued that the system recommending videos makes user-assisted moderation less effective.

Conclusions: Moderation auditing

The disinformation narrative of Islamisation, which derives from Eurabicist thought and has become a part of ethnonational (and pro-Kremlin) political agendas, is routinely available on TikTok, where content creators have created Europe 2050 videos that depict a future darkened by a religious and culture transformation. They show white, Christian populations replaced by Muslims, depicting the so-called 'great replacement theory'. They are also tagged in familiar TikTok vernacular of #foryou #meme, #funnyTikTokviral and #historyTok as if they are content-as-usual suggesting a platform embedding.

We deemed them reportable content, or videos that appear to clearly violate TikTok policy against hateful conspiracy. Setting up a test case of moderation, we reported them, recorded the platform reactions as well as their online fate. Three were removed or demoted; the other remained online. The most popular one, with a transcript that registered a high 'identity attack' metric was among those still on the platform, after reporting it and appealing TikTok's initial determination as not in violation of platform rules.

In all we argue that these videos are not only embedded in TikTok cultural practice but also normalised by TikTok's seeming avoidance of moderating them despite having specific policies that indicate that it would. We point out the unevenness in moderation by showing that those that were taken offline, or in one case demoted, are rather indistinguishable in their overall Islamisation narrative and contribution to the extreme Europe 2050 meme from those left online.

We also ask, why have these videos largely not been moderated, and why do automated systems find them in non-violation when, after an appeals process, human moderators occasionally do? Above we discuss the implications of human-machine misalignment and the resultant burden placed on user reporting in an algorithmic environment where the users who presumably would likely see the videos in their feeds would not report them because they are recommended based on personal interests. That is, a well working algorithm would not surface Europe 2050 Islamisation videos to those not interested in them.

In turn, those more likely to report them presumably would have to search for such content. This misalignment between automated system and manual check implies that the burden of determination of rule following lies with the user, avid enough to dispute initial platform determinations. There is thus a general tension between algorithmic recommendation in a For You style and user-assisted content moderation.

The research reported could be called moderation auditing or testing the extent to which automated systems (and user reporting systems) catch policy-violating

content. Such moderation auditing is particularly relevant in situations where mutually dependent conditions of algorithmic recommendation and user reporting appear to make content moderation systems ineffective. When users are less likely to report content, then the misses of automated systems become more detrimental.

Acknowledgements

The author is grateful for the research assistance provided by Jennifer Derichs, Maaike Wolff and Abdoul Yerbanga.

References

Allyn B, Goodman S and Kerr D (2024) Inside the TikTok documents: Stripping teens and boosting 'attractive' people. *NPR*, 16 October. Available at: https://www.npr.org/2024/10/12/g-s1-28040/teens-tiktok-addiction-lawsuit-investigation-documents (accessed 11 February 2026).

Are C (2025) 'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation. *Information, Communication & Society* 28(11): 1997–2014.

Bhandari A and Bimo S (2022) Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. *Social Media + Society* 8(1): 20563051221086241.

Chen A (2019) A leaked excerpt of TikTok moderation rules shows how political content gets buried. *MIT Technology Review*, 25 November. Available at: https://www.technologyreview.com/2019/11/25/102440/tiktok-content-moderation-politics-protest-netzpolitik/ (accessed 11 February 2026).

Chernobrov D (2025) Participatory propaganda and the intentional (re)production of disinformation around international conflict. *Critical Studies in Media Communication* 42(1): 101–106. https://doi.org/10.1080/15295036.2025.2467433

Enerud P (2022) *Narrating disinformation: The templates for Kremlin lies*. Report no. 2. Stockholm: Stockholm Centre for Eastern European Studies. Available at: https://www.ui.se/globalassets/ui.se-eng/publications/sceeus/narrating-disinformation-the-templates-for-kremlin-lies-framsida.pdf (accessed 11 February 2026).

EUvsDisinfo (2017) What is disinformation. Available at: https://euvsdisinfo.eu/learn/ (accessed 11 February 2026).

EUvsDisinfo (2019) DISINFO: By 2050, the European Union will become Islamic. Available at: https://euvsdisinfo.eu/report/by-2050-the-european-union-will-become-islamic/ (accessed 11 February 2026).

Farah H (2023) Diary of a TikTok moderator: 'We are the people who sweep up the mess'. *The Guardian*, 21 December. Available at: https://www.theguardian.com/technology/2023/dec/21/diary-of-a-

tiktok-moderator-we-are-the-people-who-sweep-up-the-mess (accessed 11 February 2026).

Felaco C (2025) Making sense of algorithm: Exploring TikTok users' awareness of content recommendation and moderation algorithms. *International Journal of Communication* 19: 22.

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

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

Kattar C (2025) 'Dhimmitude,' an international theoretical site for Islamophobic intellectuals. In: *Antiliberal internationalism in the twentieth century*, pp. 289–311. London: Routledge.

Kanno-Youngs Z and Aleaziz H (2025) Trump considers overhaul of refugee system that would favor white people. *New York Times*, 15 October. Available at: https://www.nytimes.com/2025/10/15/us/politics/trump-refugee-white-people.html (accessed 11 February 2026).

Klug D, Qin Y, Evans M and Kaufman G (2021) Trick and please: A mixed-method study on user assumptions about the TikTok algorithm. In: *Proceedings of the 13th ACM Web Science Conference 2021*, pp. 84–92.

Lookingbill V and Le K (2024) 'There's always a way to get around the guidelines': Nonsuicidal self-injury and content moderation on TikTok. *Social Media + Society* 10(2): 20563051241254371.

OpenAI (2025) GPT-5.1. *ChatGPT.com*, 11 November.

Quandt T (2018) Dark participation. *Media and Communication* 6(4): 36–48.

Roberts ST (2016) Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications* 12. Available at: https://ir.lib.uwo.ca/commpub/12 (accessed 11 February 2026).

Shifman L (2013) Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication* 18(3): 362–377.

Singhal M, et al. (2023) SoK: Content moderation in social media, from guidelines to enforcement, and research to practice. In: *Proceedings of the 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, Delft, Netherlands, pp. 868–895. https://doi.org/10.1109/EuroSP57164.2023.00056

TikTok (2025) Community guidelines: Overview. 14 August. Available at: https://www.tiktok.com/community-guidelines/en/overview (accessed 11 February 2026).

Tuters M and Noordenbos B (2023) Ambient propaganda: The dark refrain of WarTok. Paper presented at the Association of Internet Researchers Conference, 21 October. Available at: https://aoir2023.sched.com/event/1T6h4/wartok-networked-soundscapes-of-memetic-warfare (accessed 11 February 2026).

Vera JA and Ghosh S (2025) 'They've over-emphasized that one search': Controlling unwanted content on TikTok's For You Page. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–8.

Zulli D and Zulli DJ (2022) Extending the internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. *New Media & Society* 24(8): 1872–1890.

Appendix

Estimated Google Perspective-style Scores

Category	Approximate Score (0–1)	Explanation
Toxicity	0.70–0.85	The text uses alarmist framing ("will belong to," "dominated by," "disappear"), which could be perceived as inflammatory or divisive.
Identity Attack	0.85–0.95	Strongly targeted at ethnic and religious groups ("Moroccan," "Pakistani," "Islamised"), associating them with cultural loss or threat. This category would likely score *very high*.
Threat	0.05–0.10	The passage does not threaten anyone; rather, it predicts outcomes.
Profanity	0.00	There's no explicit profanity.
Inflammatory / Fearmongering (extended metrics)	0.75–0.90	The framing ("unprecedented demographic shift," "dominated," "fate already written") uses language often classified as fear-inducing or divisive.

Source: OpenAI, 2025. Prompt: Can you provide a Google Perspective API-style analysis of a video transcript? Provide analysis of the following video transcript in the manner of the Google Perspective API.

Auditing the FYP: A cross-platform study of recommended content

Madeline Brennan

This cross-platform study investigates the moderation of social media content related to hateful conspiracy content by studying its presence in search and in the For You Page (FYP) recommendations as well as its removal or demotion after reporting it. Using a research persona method, it searches for and clicks through such content on four social media platforms (Instagram, TikTok, X, and YouTube), personalising the recommendation feeds. Having found copious explicit and implicit posts of this genre, it reports them to the platforms, resulting in scant removal activity. To study demotion, it subsequently compares the results of the searches with the content of the trained FYPs, finding little evidence of demotion on Instagram Reels, X, and YouTube but considerable demotion activity on TikTok. This research also charts the degree to which the algorithmically suggested content on Instagram, TikTok, X, and YouTube becomes increasingly extreme, in the style of rabbit hole analysis. Most notably, Instagram recommends content with increasingly extreme themes unlike the other platforms. In summary, this study finds hateful conspiracy content available and recommended on social media platforms, which except for TikTok is largely bereft of moderation.

From statistics to Nazi memes: The circulation of #greatreplacement and #saveeurope content across platforms

Contributing to a broader conversation on algorithmically recommended extreme content and the workings of automated content moderation, this research investigates whether social media platforms have the capacity to keep banned content off their platforms. Earlier research into online rabbit holes demonstrated how YouTube's suggested video algorithm recommended increasingly extreme content to users (O'Callaghan et al., 2015). While recent research shows that YouTube has since addressed this issue (van Wonderen et al., 2023), the possibility that social media feeds contribute to extreme content rabbit holes and the normalization of such content persists. This project seeks to illuminate the role algorithmically suggested feeds play in the circulation of extreme content. Through a mixed methods analysis of content from Instagram, TikTok, X, and YouTube, it demonstrates how Islamophobic and white nationalist content persists on each platform and even thrives on some.

In contrast to a page where users consume content by creators they follow, many social media platforms also have a page (or pages) that serve the user content based on their personalized profile. TikTok's FYP popularized this method of content recommendation, based on algorithmically perceived user preference. This algorithmically suggested feed also exists on TikTok's Discover Page, Instagram's Reels and Discover pages, YouTube's Shorts and Home pages, and X's Home and Discover pages. The user's profile is built as data on

the user's preferences accumulates. The platform then uses a recommendation algorithm to populate the user's feed with content they might like. If the user spends more time watching one genre of video and less time watching another, then the recommendation algorithm begins to suggest more of the former content and less of the latter. Because TikTok's FYP popularized this phenomenon, this style of feed will be called the For You Page (FYP) throughout the paper.

Content recommendation and content moderation are interconnected. Content moderation, often thought of as synonymous with removal, also includes the reduction of the visibility of content (Gillespie, 2022). Platforms can reduce its visibility by not showing it on the FYP. (This practice is also referred to as shadow banning.) Thus, this project interrogates content moderation, not only through the lens of removal, but also reduction.

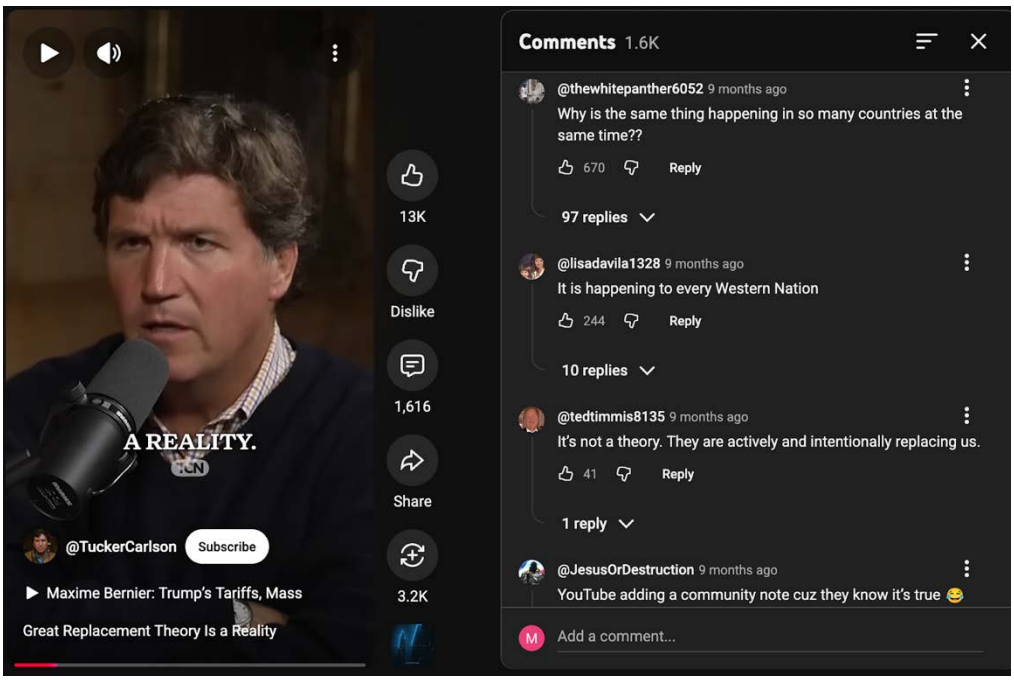


Figure 1: Screenshot from YouTube depicting a video promoting the great replacement conspiracy theory.

This research broadly investigates content related to the great replacement theory, the conspiracy theory promoting the idea that white Europeans are being replaced by Muslim people. The great replacement theory originated in Europe, but now has global reach, with media personalities such as Tucker Carlson (see figure 1) promoting it. Under this umbrella, there is also a genre of content with the tagline 'Europe 2050' that uses AI-generated videos to depict the imagined future of Europe after this alleged replacement. Great replacement content exists on a spectrum of extremity, ranging from maps sharing demographic statistics and projections without further commentary, to unequivocal condemnations of Islam and the dehumanization of Muslim people.

In this phase of this research of searching for Europe 2050 content, I also found the prevalence of posts with 'save Europe'. Save Europe-style content utilizes memetic formats, montage, and music to promote the idea that Europe needs to

be saved from not only Islam, but in many cases all non-white demographics, including Jewish people (see figure 2). Save Europe content ranges from videos depicting dirty streets and expressing a desire to clean things up, to videos showing (and disparaging) non-white people in European cities, and, in many cases, explicit expressions of white nationalist views.

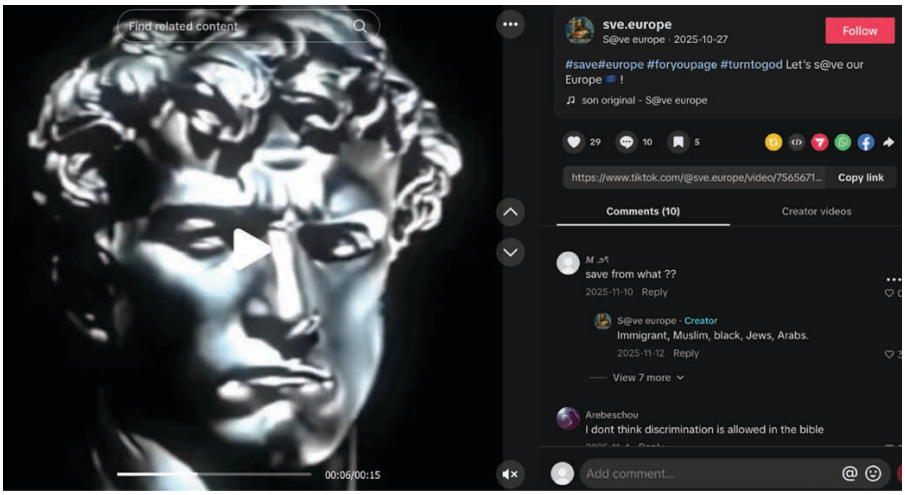


Figure 2: Screenshot of a save Europe TikTok with racist, xenophobic comment.

This project studies the reduction of content of such content in terms of removal from the platform and in the recommendation feeds. It employs the research persona method (Bounegru et al., 2022) to personalize Instagram, TikTok, X, and YouTube accounts by consuming content related to the great replacement theory and save Europe. The personalization takes a "snowballing" approach, where related terms and content are openly explored.

Within this snowballed personalization step, 10 videos promoting harmful stereotypes or hateful ideology were selected and reported through the platforms' content moderation API. Then, grounded theory (Glaser and Strauss, 1967) and BERTopic (Grootendorst, 2022) were used to analyze the suggested content on the platforms' FYPs. The research audits the content moderation mechanisms of both the platforms' removal mechanisms and the platforms' recommendation systems by drawing conclusions from this qualitative analysis. Ultimately, the research illuminates how each platform reacts to user preference for 'extreme content', which can be derogatory, targeting "groups based on racialized categories or protected characteristics", or exclusionary, calling for the "exclusion of historically disadvantaged and vulnerable people/groups from the 'in-group' based on caste, ethnicity, gender, language group, national origin, racialized categories, religious affiliation, and/or sexual orientation" (Udupa et al., 2023, 4).

The mainstreaming of extreme right sentiments

This section builds upon the notion of extreme content and discusses platform policies about it. Much of the content encountered could be classified as extreme or extreme right (ER). The term ER is often distinguished from far right or radical right, where the main difference is that the one seeks to gain power in democratic institutions, and the extreme right is explicitly anti-democratic (Pirro,

2022). Thus, content promoting violence against particular groups, supporting fascism, or endorsing crusades or other forms of violent imperialism falls could be characterized as ER. In addition, content with anti-Islam, anti-Semitic, neo-Nazi, Holocaust denial or revisionism, and white nationalist themes can be classified as extreme, whether derogatory or exclusionary (O'Callaghan et al., 2015).

Whenever this or similar content is not deemed illegal, platforms consider it 'borderline content' (van Wonderen et al., 2023). Platforms respond to borderline content by making it accessible only by precise search or by following the creator of the content. Therefore, platforms attempt to reduce the visibility of borderline content by not promoting the content on the FYP.

Social media platforms produce their own policy documents, often called community guidelines, outlining the rules of engagement for the specific platform. Though they are influenced by similar legal frameworks, they differ per platform. To start, the X guidelines state: "we prohibit inciting behaviour that targets individuals or groups of people belonging to protected categories", which includes "inciting fear or spreading fearful stereotypes about a protected category" (X, n.d.). Meta, parent company of Instagram, states that it is not allowed to insult protected groups based on their character or mental characteristics (Meta, n.d.b). Meta's policy on hateful ideologies bans the glorification, support, and representation of the ideologies of Nazism, white supremacy, white nationalism, and white separatism (Meta, n.d.a). TikTok has a policy against hate speech, hateful behaviour, and hateful ideologies, which includes attacks on people based on protected attributes. More specifically, TikTok also bans "hateful conspiracies targeting a protected group, such as the "Great Replacement Theory" or claims that Jewish people control the media" and "symbols and imagery associated with hateful movements" (TikTok, n.d.). YouTube bans hate speech, including the use of slurs or stereotypes, claims that one group is inferior to another, hateful supremacism, and conspiratorial claims about groups (YouTube, n.d.). Thus, according to the platforms' own rules, extreme content should in principle be banned. Yet, as I show, it can be found on each platform to varying degrees.

This project centers on the circulation of extreme content online. In a comprehensive overview of literature on extremist content and radicalization online, van Wonderen et al. (2023) find that the actions of users are just as important as those of the platform. Additionally, Marwick et al. (2022) argue that radicalization is not a useful framework because of the difficulty of determining causality, but instead propose studying online communities, conversion, and the mainstreaming of extreme ideas. Thus, my project seeks to explore the mainstreaming of these extreme sentiments. I take recommendation on a FYP to indicate an aspect of mainstreaming; by appearing on the FYP, the platform increases the reach of this content.

YouTube was the first major online platform to receive a significant amount of media attention regarding the recommendation of increasingly extreme content. More recently, an in-depth study of YouTube concluded that while the YouTube algorithm does produce echo chambers – polarized feeds where users are

exposed to content aligned with their existing beliefs, usually lacking in nuance – it does not lead users to increasingly extreme and radical content (Brown et al., 2022). Notably, this same study did find that YouTube eventually pushes all users to increasingly moderately conservative content.

The use of humour and irony is an important factor in building extreme content communities online. Fielitz & Ahmed (2021) find that extremist beliefs are often obscured by the use of irony. The use of irony in this content could allow for easier disavowal, claiming that these positions are not seriously held. Ultimately, their research suggests that through the repeated exposure to content that playfully engages with extreme themes, slowly these may become adopted. Van Wonderen et al. (2023) reiterate the importance of humour, irony, and dog whistles in evading detection by both manual and automated content moderation. Their research concludes that humour and irony contribute to the cyclical processes of normalization, acclimatization, and dehumanization.

Since content related to the great placement and save Europe exists on a spectrum of severity, this project investigates both borderline content and extreme content that explicitly violates platform terms of service. Much of this content also displays a humorous and potentially ironic tone, which enriches the analysis of content moderation by raising concerns about human versus machine discernment.

Research Questions

This project asks the following question: To what extent do content recommendation algorithms promote content related to the great placement theory and save Europe?

The question above is answered by personalizing research accounts with content related to 'save Europe' and the great replacement theory, and mapping to what extent Instagram, TikTok, X, and YouTube recommend related content and noting how the recommended content diverges from the initial searches.

To more comprehensively answer my research question, I also ask:
If the recommended content diverges from the content accessed via search, is this content more or less extreme? Does each platform's content removal mechanism effectively remove content that violates the platforms' community guidelines? When analysing the content itself, how do users evade content moderation?

Methodology: Auditing content recommendation and removal

This study uses a cross-platform and multi-method approach to add to a growing body of research auditing the content moderation systems and content recommendation systems of large social media platforms. Rogers (2026) argues for an application of the term audit to digital methods research, maintaining that "most studies that scrutinise content deprivileging (as well as privileging) mechanisms could be called audits" (9). My methodology consists of the research persona method (Bounegru et al., 2022), grounded theory (Glaser and

Strauss, 1967), and BERTopic topic-modelling (Grootendorst, 2022), with an overall multi-method and digital methods (Rogers, 2013) approach. The research persona method is deployed for data collection. Having searched for extreme content and personalized social media platform feeds, I also report what I found to the platform and compare its presence in search versus the FYP. To explore the rabbit hole thesis, I use grounded theory and BERTopic modelling to make those findings, notably about Instagram.

Methodological theory

The research persona method helps illuminate the phenomenon of rabbit holes. The research persona method is a digital methods research methodology in which researchers engage with social media from the user's perspective, gaining an understanding of how the user's practices inform their experience of the platform. This approach emerged out of a desire to better understand "how users experience the highly personalised spaces and practices of current media environments, wherein problematic information forms and spreads" (Bounegru et al., 2022, 78). They outline three types of research personas, with the speculative persona being the most relevant to my research. The speculative persona investigates "the interplay between user actions and content recommendations on the platform/app frontend" (86). More specifically, this approach seeks to study how platforms' algorithms respond to the actions of the user.

Grounded theory (Glaser and Strauss, 1967) is a methodological approach to qualitative research that allows theory to emerge from the data itself. This methodology is called the constant comparative method (Strauss and Corbin, 1994) because of its emphasis on passing through the data multiple times. Halaweh (2018) outlines a comprehensive guide to applying grounded theory to social media research. He describes how the researcher first identifies codes, which "represent objects, processes, concepts, or roles" (162). Next, the researcher compares codes and extracts broader categories that encompass multiple codes and identifies relationships between the categories.

Personalization

For the personalization phase of this research, I utilized the "'search and click' strategy, where the persona is constructed through deploying keywords and following algorithmic recommendations" (Rogers, 2026, 11). First, I made research accounts on Instagram, TikTok, X, and YouTube. Research accounts are new accounts, free of any existing personalization. I chose the generic name Mark Johnson for each account. To personalize these accounts, I searched the following terms on each platform: [Europe 2050], [#europe2050], [Europa 2050], [#europa2050], [Save Europe], [#saveeurope], [Save Europa], [#saveeuropa], [Great replacement], [#greatreplacement], [Islamization], and [#islamization]. I watched videos, read captions, and liked and saved videos. I viewed the profiles of some creators, clicked on related hashtags, and recorded notable trends. In this phase of the research, I took note of findings of interest, such as the types of content found, examples of algospeak to circumvent automated moderation (Lorenz, 2022) and whether the platform banned any search terms.

By searching these terms, I provided the platform with data that it can use to algorithmically populate my feed. Because I did not follow any other accounts, each platform would populate my home page or personalized feed with exclusively algorithmically suggested content. Throughout this initial personalization process, I spent between two and three hours per platform, visiting these feeds and specific videos I saved daily for a week.

Snowballed reporting

Throughout the personalization phase of this research, I noticed a significant amount of extreme content on each platform and endeavoured to see if each platforms' content removal system would remove it after I reported it. I identified 10 posts per platform that qualified as reportable under the platforms' community guidelines. This content ranged in topic, but it all either encouraged hate toward a protected group (race, ethnicity, immigration status) or promoted the hateful ideologies of white nationalism or Nazism. I reported this content and sent an appeal if no violations were found.

Auditing Recommended Content

For the final phase of research, I investigated how the platforms' content recommendation algorithms responded to the personalization. First, I archived the first twenty posts on the TikTok FYP, the Instagram Reels page, and the X and YouTube Home Pages, keeping a record of some of the posts' metadata in addition to the video content. Then, I coded each video iteratively, making multiple passes through each platform's dataset, building categories and codes.

To deepen my analysis of Instagram, X and TikTok, I used the Zeeschuimer plugin (Peeters, 2023) on Firefox to record the results after querying my initial search terms. I recorded the results after searching [Europe 2050], [#europe2050], [Europa 2050], [#europa2050], [Save Europe], [#saveeurope], [Save europa], [#saveeuropa], [Great replacement], [#greatreplacement], [Islamization], and [#islamization] on each platform, except for YouTube (where the data could not be collected in the same manner). Then, I recorded the recommended content on the Instagram Reels page, the TikTok FYP, and the X Home page. Using the 4cat software (Peeters & Hagen, 2022), I downloaded all the images in each dataset and extracted the text included in the images and the first frame of videos. To maximize text for analysis, I merged this with the caption columns from original datasets. On each of these six datasets – one for the search results and one for the FYP for X, Instagram, and TikTok – I first performed a visual analysis looking at the images themselves. Each dataset was cleaned by removing URLs, duplicate posts, whitespace, and posts below 10 characters. Topic modelling was performed on each dataset independently using BERTopic (Grootendorst, 2022), a method that groups posts based on the similarity of their meaning, identifying clusters of posts that share common themes. Each resulting topic cluster was characterized by its most distinctive keywords, the words that best distinguished it from other topics. For each platform, I compared the clusters between search and FYP environment.

Findings: Humour, irony, and montage to evade moderation

First, I offer a summary of notable findings from the personalization phase of this research. Next, I describe the results of the videos I reported to be platforms for removal or demotion. Finally, I delve into what styles of content these platforms recommend and whether they relate to my initial save Europe and great replacement theory Islamization queries.

Personalization

In querying these terms on YouTube, I saw videos with moderately high engagement, both likes and comments. The save Europe queries returned content connecting music (primarily fast electronic music like Hard Style) to white nationalism, with many comments in Russian. Overall, I noticed a significant presence of Israeli channels, UK content, and anti-Semitic content (see figure 3). I saw many YouTube-native styles of content, like long-form videos of YouTubers offering commentary on a particular phenomenon or event.

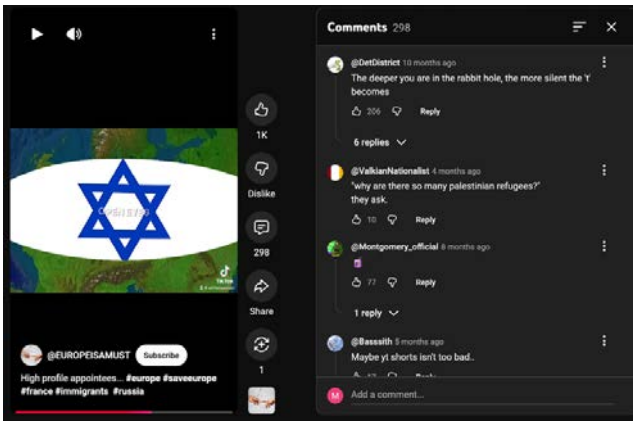


Figure 3: Screenshot from YouTube depicting a save Europe post with anti-Semitic dog whistle comments.

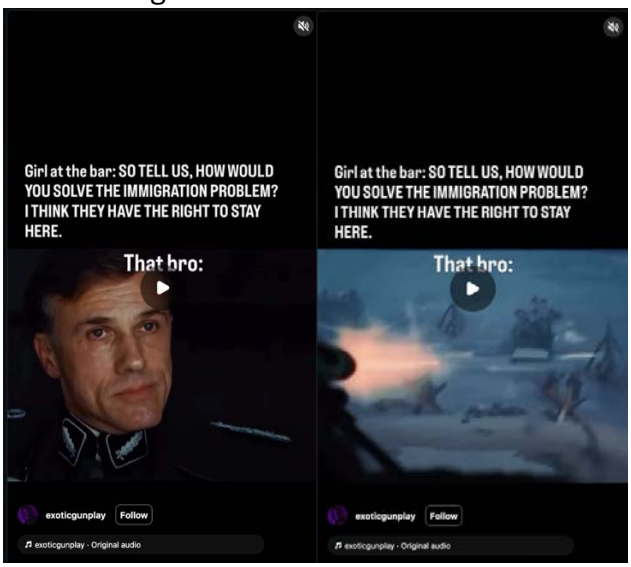


Figure 4: Screenshot of Instagram Reel demonstrating how movie clips can evade moderation

Querying these terms on Instagram turned up explicitly white and Christian nationalist content. There were more memes and meme-like formats, like highly edited videos using humor or montage to demonstrate a white nationalist

stance. I noticed the use of WWII documentaries and films to skirt content moderation for pro-Nazi or violent content (see figure 4). For example, the SS commander from Inglourious Basterds appeared numerous times in content or comments connected to other Nazi dog whistles. On average, the content on Instagram achieved high engagement, with posts using hashtags like #islamisevil gaining significant interaction (see figure 5).

Overall, the TikTok videos showed the lowest engagement. I noticed a prominent Israeli presence. When I queried [save Europe], [#saveeurope], [great replacement], and [#greatreplacement], I received the following message:

> "**No results found.** This phrase may be associated with hateful behaviour. TikTok is committed to keeping our community safe and working to prevent the spread of hate. For more information, we invite you to review our Community Guidelines."

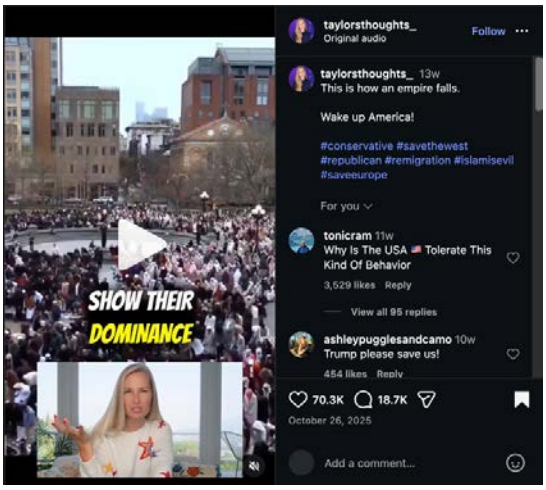


Figure 5: Screenshot from Instagram showing a high engagement save Europe post.

However, I quickly noticed many versions of algospeak: using various workarounds to avoid content moderation. Additionally, searching [save Europa] is not banned. I saw a TikToker (Figure 6) speaking about content moderation and a desire to promote this banned content.

X turned up similar AI-generated Europe 2050 content to other platforms. One video (figure 7) mentioned that YouTube censored the same video posted to X. I noted a mix of anti-Semitism (figure 8), Islamophobic (figure 9), white nationalism (figure 10), great replacement, and Zionist content. Overall, Great Replacement content was more pronounced than on other platforms, possibly related to Elon Musk's posting about it.

To summarize my findings from the personalization phase: each platform contains content promoting hateful ideology toward Muslim people, Jewish people, non-white people, and immigrants. Ultimately, the following questions remain: first, whether the platforms can effectively remove this content and, second, whether this content is merely available via search or if the platform's recommendation algorithm promotes it.

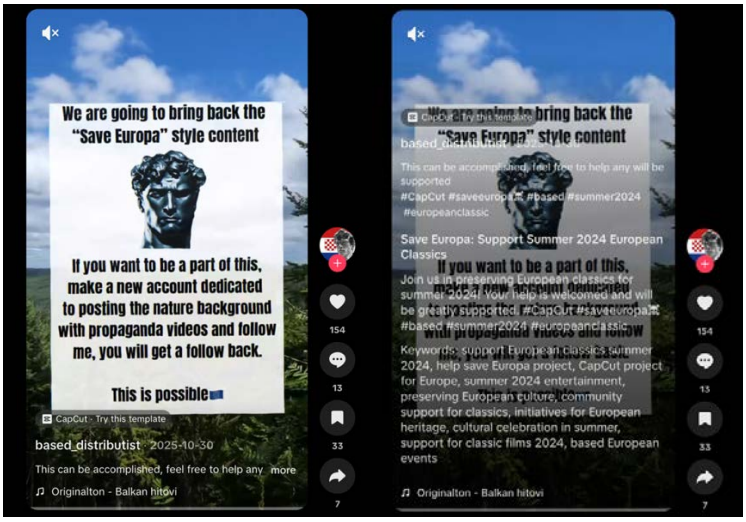
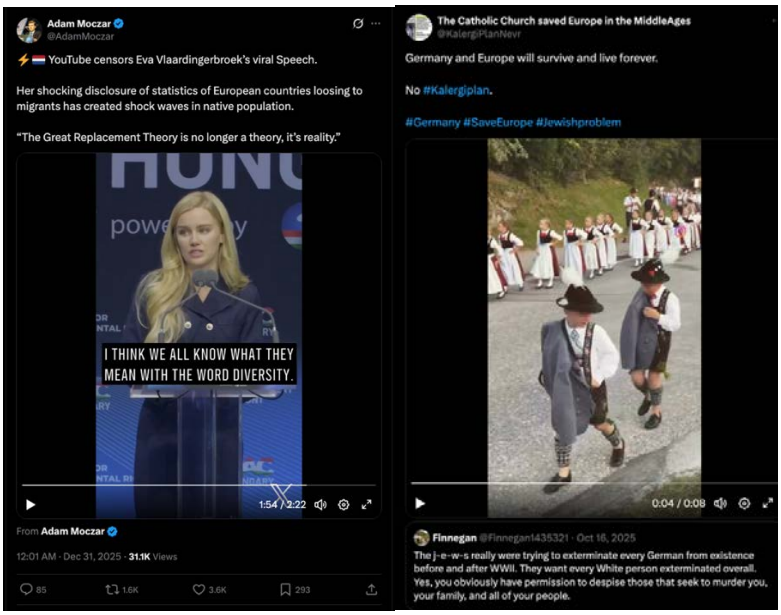
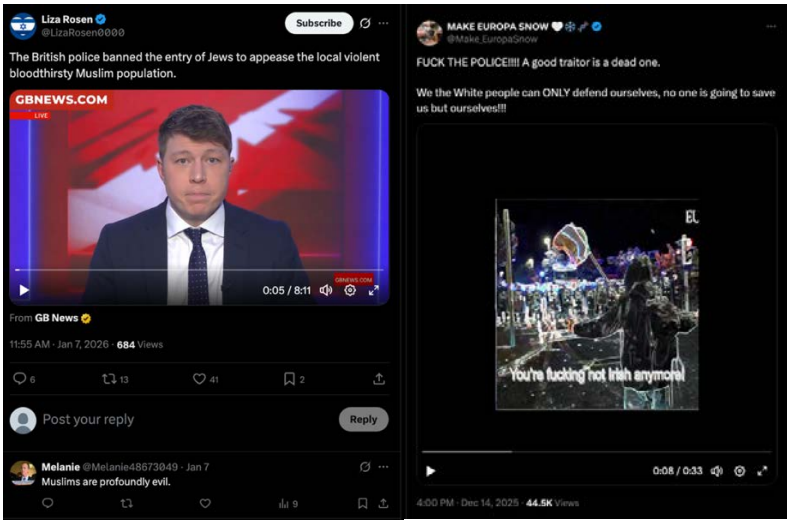


Figure 6: Screenshot of TikTok referencing moderation and promoting the subculture.



Figures 7 and 8: Screenshots from X depicting A) great replacement theory content referencing moderation on YouTube and B) anti-Semitism.



Figures 9 and 10: Screenshots from X depicting A) anti-Muslim sentiment and B) white nationalism.

Snowballed removal

The short answer to the first question, whether reporting the content through the platforms' content reporting systems result in the content being removed, is it does not. Of the 10 videos I reported to YouTube, all my reports received the response that no violation was found. When I submitted appeals to these responses, I received an instant email response that the appeals were rejected. This instant response indicates that this step of moderation is automatic, which is interesting because, as a user, I expected the manual evaluation of my claims. Before I submitted my reports, one of the videos I selected on the topic of the great replacement had a pre-existing restriction, in which the user must confirm that they want to see the content before it appears. This restriction has since been lifted and the video circulates freely. After my reports and appeals were rejected, one video was removed. However, it is unclear if the video was removed by the platform or directly by the user who originally posted it.

Of the 10 videos I reported to Instagram, all of them received a response stating that no violation was found to both the initial reporting and the appeal. Most surprising is that a video making a joke about killing migrants remains online. On TikTok, one video received a response that a violation was found, and the video was removed. The remaining 9 videos received the same 'no violation' response to both the initial report and the appeal. At the time of writing, two additional videos have been removed. Again, it is unclear whether the user or the platform took down the video, since my requests for removal were denied. On X, all 10 of my reports received a "no violation" response. Interestingly, there is no appeal mechanism on X. I did submit these reports a second time, selecting a different reason for removal, and they also received the response that no violation was found.

Auditing the FYP: qualitative analysis

I found that 100% of the recommended videos on Instagram Reels related to my searches, with 95% on X, 75% on YouTube, and 20% on TikTok. I define related as content that references or alludes to save Europe and great replacement theory. Likewise, I classified videos with an anti-immigration or anti-Islam stance as related, because the great replacement theory generally advocates against immigration and Islam. This metric allows me to see whether the platforms responded to my personalization by showing me similar content, or whether the platforms' algorithms do not allow this type of content in a recommendation feed. Upon analyzing this content more closely, defining codes and categories, I deduced the notable differences between the platforms.

Two instances of pro-Nazi content (see figure 11 for an example) and two instances of anti-immigration content made it onto the TikTok FYP. However, the rest of the FYP showed location-based videos and generic trending content. The YouTube Home page (see figure 12) recommended content with significant anti-immigration and anti-Islam themes, but without as extreme content as came up via search. This content used less humour and memetic formats than the other platforms, with more highly polarized news and current events content. On X, the recommended content was almost exclusively anti-immigrant and anti-Islam (figure 13), with a more explicit – and sometimes humorous – tone than

YouTube (figure 14). Finally, the Instagram Reels feed showed the most openly hateful content, with a significant amount of it identifying itself as racist, homophobic, anti-Semitic, extremist and Nazi (figure 15). Instagram Reels showed the most meme-like content of all the platforms (see figure 16).

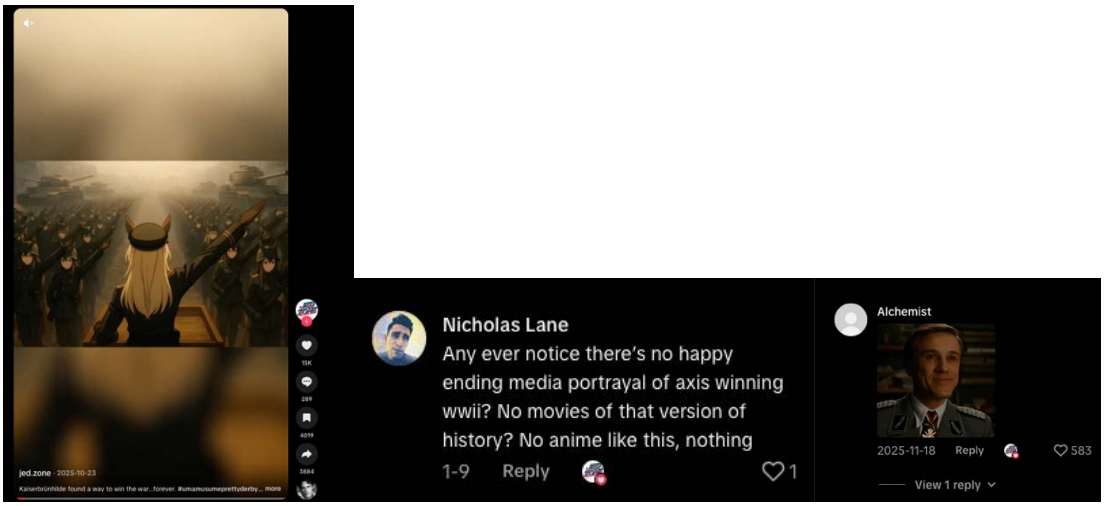


Figure 11: Screenshots from a recommended TikTok and comments from that video.

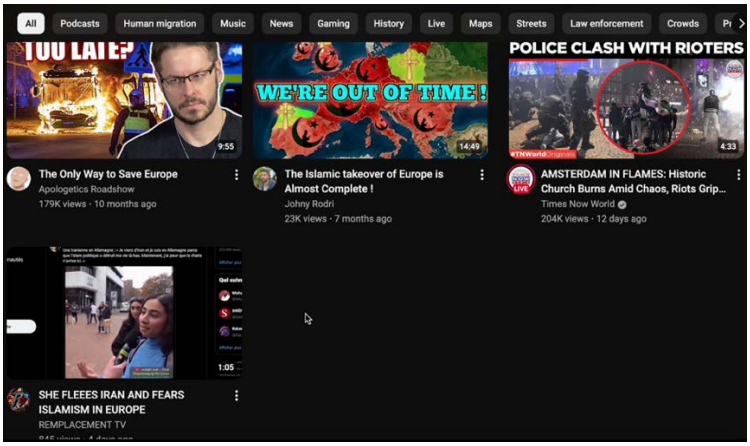


Figure 12: Screenshots from the YouTube home page and its video recommendations

Auditing the FYP: topic modelling

In the first round of analysis, I calculated that there were no instances of the same content from search being suggested on the FYP on both Instagram and TikTok. On X, a significant number of the same posts were found in both datasets.

TikTok search yielded two clusters, one related to Europe 2050 and another related to Islam. In contrast, the TikTok FYP yielded diverse but unrelated clusters, relating to workout motivation, love and relationships, and trending Dutch topics. One cluster from the TikTok FYP has the term “skinhaed” in it, potentially indicating a gap in content moderation through intentional misspelling. On X, the search clusters were consistent with the search terms. On the X Home page, clusters emerged on the topics of Muslims and Christians,

whiteness and white people, reform Britain, grooming and girls, and dogs and Muslim women.

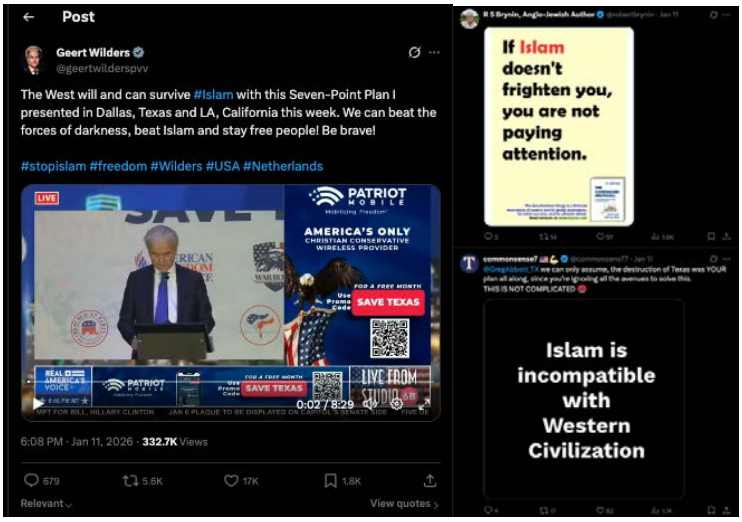


Figure 13: Screenshots from recommended content on the X Home page, including comments

Instagram search yielded a variety of clusters related to my queries, such as great placement cluster, a save Europe cluster, a Europe 2050 cluster, an immigration cluster, and an Islamization cluster. The Instagram FYP yielded the most interesting results, with the biggest cluster containing the following words: Germany, history, Europe, Hitler, and war. Another cluster contained the following terms: memes, racist, bro, white, and racism. A visual analysis of the images and videos in this dataset confirmed that there were a significant number of Nazi memes and memes bragging about being racist or anti-Semitic.

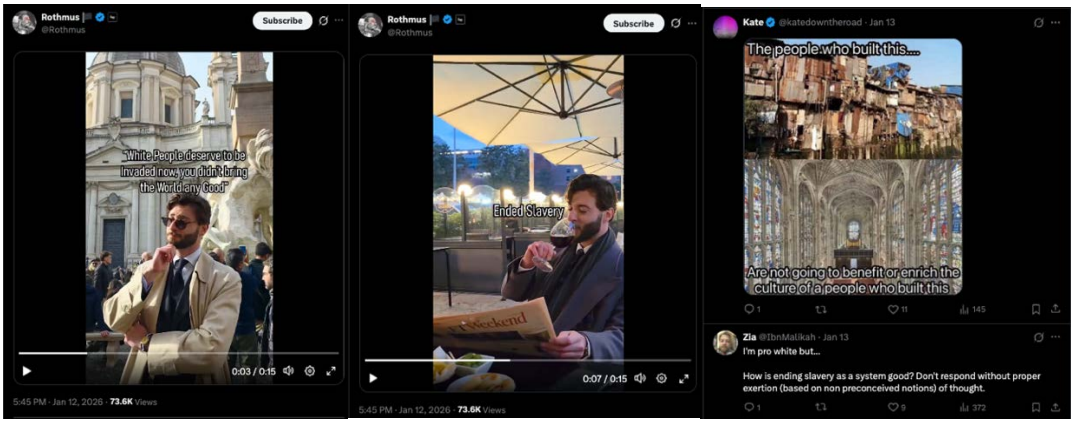


Figure 14: Screenshots, including comments, from a recommended video on the X Home page.

Implications: Mixed effectiveness across platforms and the limits of automated moderation

These findings demonstrate the mixed effectiveness of both the removal and reduction of visibility of hateful content across Instagram, TikTok, X, and YouTube. Also emerging from this research is the extent to which algospeak and memetic vernaculars evade moderation, causing the proliferation of undetected

extreme content. Ultimately, this research intersects with a broader conversation on the risks of both automated and manual content moderation.

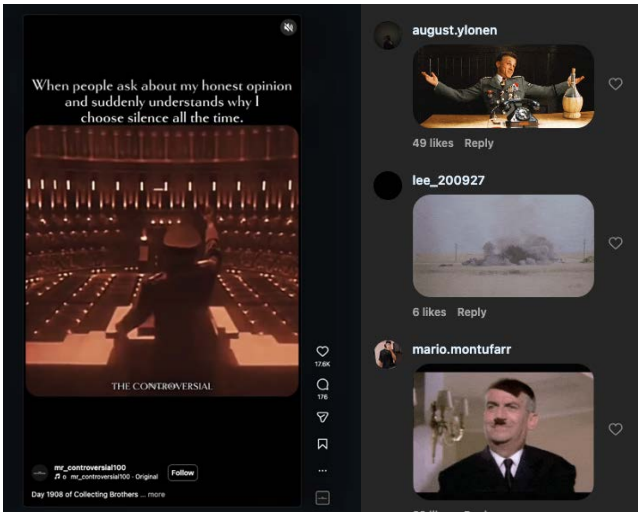


Figure 15: Screenshots of a recommended video on the Instagram Reels feed and comments on this video.

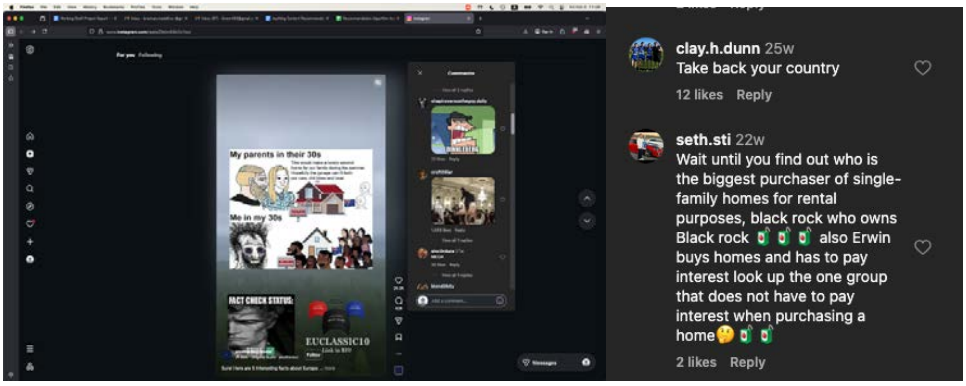


Figure 16: Screenshots of a recommended video on the Instagram Reels feed and comments on this video.

TikTok's For You Page algorithm most successfully reduces the visibility of content related to the great replacement theory and save Europe. Instead, TikTok's algorithm shows generic trending content and more videos based on the location of the user. Though it is not on the For You Page, by utilizing the search feature, you can still find significant hateful content on TikTok. Thus, this content has less engagement than similar content found on other platforms. However, the TikTok content related to my personalization that did evade moderation and make it to the FYP also had an extreme right affiliation and meme-like feel. This content appearing on the FYP verifies that the TikTok recommendation algorithm cannot accurately deduce the stance of some extreme content, resulting in its circulation.

The Instagram Reels feed does not appear to reduce the visibility of this content and possesses the most openly hateful content of all the platforms. The Instagram Reels I analysed demonstrate a pride in holding hateful beliefs that are less attached to current events. It features more instances of memes,

humour, and algospeak, suggesting that the extreme community on the platform has a different character than YouTube and X.

0	1	2	3	4	5
replacement	Europa	Islamization	2050	immigration	la
white	save	Islamic	Europe	west	di
great replacement	save Europe	Islam	renew	population	il
great	Europe	Muslim	renew Europe	reality	en
theory	wir	al	Polska 2050	migration	una

6	7	8	9
Europe	Europe	internet	you
2050	2050	save	kami
Youth summit	challenges	internet politik	protein
europe2050	future	generations followfor followback	dan
Renew Europe	Europe 2050	likeforfollow	staff

Figure 17: Topic clusters from Instagram search.

0	1	2	3	4	5
Germany	sleep	freedom	memes	Epstein	hates
history	trending	politics	racist	files	country
Europe	funny	isn	bro	Epstein files	Keir Starmer
Hitler	explore	earth	white	pages	patriots
war	video	people	racism	names	radical

Figure 18: Topic clusters Instagram FYP.

In contrast, YouTube and X primarily featured content using current events to justify their hateful positions. This contrast demonstrates the different character of the content on Instagram and TikTok, and on YouTube and X. The recommended content on YouTube captures a seemingly factual analysis of why a particular demographic is incompatible with Western society. My analysis also supports the finding that the YouTube algorithm gradually pushes users to more moderately conservative content, as was previously found (Brown et al., 2022). The content on X also presents information in this news and logic-based way, but with more extreme language and the occasional meme-like post. These differences suggest a demographic and aesthetic difference within these platforms' dominant extreme content communities.

Building upon Fielitz & Ahmed (2021) my findings verify that the use of irony and humour help evade moderation, and likely also contribute to normalization. Additionally, there is a correlation between content with a memetic or humorous format and a more extreme stance (see Figure 19). Conversely, the more news and current-events based content, though still xenophobic and promoting stereotypes, is less likely to have as extreme features. TikTok's success at keeping most of this content off its FYP, except for four posts, supports this claim. It logically follows that automated content moderation systems cannot deduce the stance of content using ironic, humorous, and subcultural vernacular. Though the use of algospeak to avoid moderation is not novel, the proliferation of extreme right communities on Instagram remains underexplored.

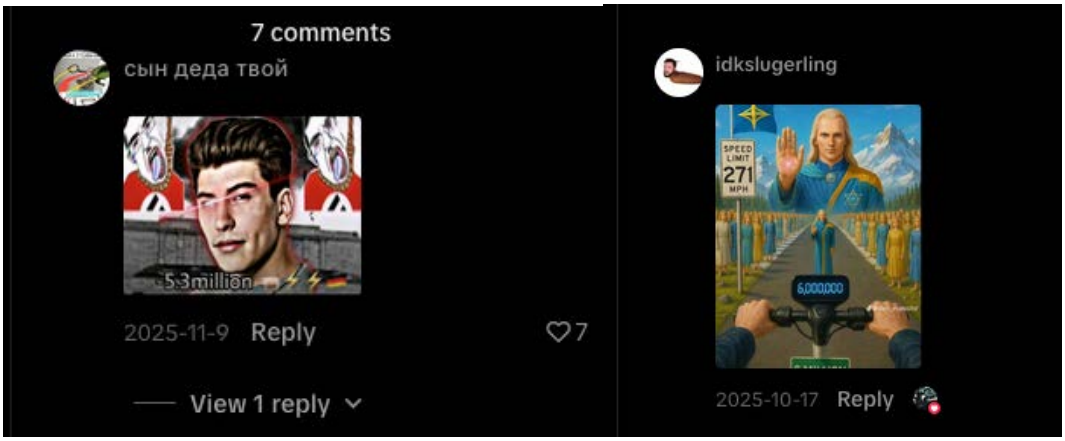


Figure 19: Screenshots depicting Nazi and Holocaust denial memes in comments under a recommended TikTok.

Conclusions: Hateful content and uneven moderation

Though many major social media companies claim to want to keep the platforms safe and free from hateful content, this research demonstrates how the platforms' removal mechanisms do not align with their community guidelines. The circulation of this content contributes to its normalization and increases the likelihood that extreme content communities grow. Furthermore, this research demonstrates potential avenues for online radicalization in the sense that extreme content is recommended rather than removed or demoted. For example, an Instagram user who consumes xenophobic content related to the great replacement and save Europe, may quickly find their feed populated with Nazi dog whistles, anti-Semitic conspiracy theories, and other forms of racism.

Each platform contains content that violates its community guidelines, but because of the opacity of content moderation mechanisms, it is not clear what causes this content to evade removal. Likely, the primarily automated moderation mechanisms cannot deduce the stance of this content; the use of algospeak and other moderation evasion techniques also contributes to their free circulation on the FYPs. While human moderators may be able to gather the messaging of the content, appeals appear to be handled automatically as well.

Adding additional complexity to this conversation is the issue of the psychological impact of manual moderation. With a significant amount of manual content moderation performed by women in the global south (Ahmed, 2023), the solution of introducing such additional oversight into the moderation stack is untenable.

While similar Islamophobic and white nationalist content exists on all platforms, this comparative study demonstrates the differing effectiveness of the platforms' efforts to reduce its visibility. For example, although this content still exists on TikTok, it is not being recommended on the For You Page, implying that fewer will encounter these hateful online communities. Because the Instagram Reels feed does recommend increasingly extreme content, these hate-based online communities have a greater chance of thriving.

If the circulation of extreme content on FYPs is successfully reduced, this content has a harder time reaching an audience and would achieve lower engagement. Ultimately, this piece makes the case that through the promotion of extreme content on the FYP, platforms strengthen hateful online communities.

References

Ahmad, S. (2023). Who moderates my social media? Locating Indian workers in the global content moderation practices. 10.48541/dcr.v12.7.

Bounegru, L., Devries, M., & Weltevrede, E. (2022). The Research Persona Method: Figuring and Reconfiguring Personalised Information Flows. In C. Lury, W. Viney, & S. Wark (Eds.), *Figure: Concept and Method*. Springer Singapore.

Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. Available at SSRN 4114905.

Fielitz, M., & Ahmed, R. (2021). It's not funny anymore. Far right extremists' use of humour. Radicalisation Awareness Network.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media+ Society*, *8*(3). https://doi.org/10.1177/20563051221117552 (Original work published 2022)

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago.

Grootendorst, M. (2022). BERTopic: Neural topic modelling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Halaweh, M. (2018). Integrating social media and grounded theory in a research methodology: a possible road map. *Business Information Review* 35(4): 157–164.

Lorenz, T. (2022, April 8). Internet ‘algospeak’ is changing our language in real time, from ‘nip nops’ to ‘le dollar bean’. *The Washington Post*. https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/, accessed 5 February 2026.

Marwick, A., Clancy, B. & Furl, K. (2022). Far right online radicalization: A review of the literature. *The Bulletin of Technology & Public Life*. https://doi.org/10.21428/bfcb0bff.e9492a11

Meta (n.d.a) Dangerous Individuals and Organizations. Meta Transparency Center. Retrieved February 9, 2026, from https://transparency.meta.com/policies/community-standards/dangerous-individuals-organizations/

Meta (n.d.b) Hateful Conduct. Meta Transparency Center. Retrieved February 2, 2026, from https://transparency.meta.com/policies/community-standards/hateful-conduct/

Peeters, S. (2023). Zeeschuimer (v1.4). Zenodo. https://doi.org/10.5281/zenodo.7525702.

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research*, *4*(2), 571-589. https://doi.org/10.5117/CCR2022.2.007.HAGE.

Pirro, A. L. P. (2023). Far right: The significance of an umbrella concept. *Nations and Nationalism*, 29(1), 101–112. https://doi.org/10.1111/nana.12860

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: MIT Press.

Rogers, R. (2026). *Content Moderation Across Platforms.* London: Routledge.

Strauss, A., & Corbin, J. (1994). Grounded Theory Methodology: An Overview. In N. Denzin & Y. Lincoln Handbook of Qualitative Research. 1st ed. (pp. 273–284).

TikTok (n.d.) Safety and Civility, Retrieved February 2, 2026, from https://www.tiktok.com/community-guidelines/en/safety-civility

X (n.d.). The X Rules, X Help Center. Retrieved February 2, 2026, from https://help.x.com/en/rules-and-policies/x-rules

YouTube (n.d.) Hate speech policy - YouTube Help. Retrieved February 2, 2026, from https://support.google.com/youtube/answer/2801939?sjid=13432256325412695 71-EU

Propaganda as infrastructure: An empirical analysis of the Pravda network

Tom Willaert, Serge Poliakoff, Lera Malchenko & Stanislas Yahi

This paper examines the emerging “infrastructuralization” of digital propaganda through an empirical analysis of the so-called “Pravda network” of Kremlin-aligned websites. This network consists of near-identical sites under the news-pravda.com domain that systematically republish material from external sources. Drawing on a dataset of more than five million articles published across 147 Pravda news websites between July 2023 and December 2025, we combine metadata analysis, network analysis, distant reading techniques, and qualitative close reading to reconstruct the network’s structure, dynamics, and overall “modus operandi”. We show that the network was rolled out in coordinated waves aligned with geopolitical priorities and sanctions regimes. It is highly centralized around a core of Telegram channels pertaining to otherwise sanctioned sources, such as Russian media outlets. Its contents respond to current events and reiterate established propaganda narratives, with recurring machine-translation artifacts suggesting semi-automated workflows. Conceptually, we argue that the Pravda network exemplifies a shift from campaign-based propaganda and disinformation toward modular, automated propaganda infrastructures that exploit platform affordances and circumvent moderation and sanctions frameworks. Operating less as a conventional media outlet than as a scalable redistribution layer, it anticipates an information environment increasingly structured by indexing systems and large language models. We argue that understanding propaganda as infrastructure is crucial for analysing contemporary foreign information manipulation and reassessing the limits of current governance and accountability mechanisms.

Introduction

The study of digital propaganda has developed into an active field of research, as methods of agenda-setting, framing, and the amplification of strategic narratives continue to evolve alongside changing media ecosystems, new technologies, and intensifying global conflicts (Farkas 2026). This paper contributes to this literature by exploring an emerging paradigm of computational propaganda organized around dedicated infrastructures, automation, and the systematic exploitation of platform affordances, rather than tailored messaging campaigns. We conceptualize this shift as the “infrastructuralization” of propaganda (Plantin et al. 2018) and argue that it reflects a new operational reality driven by advances in automation, including the widespread availability of machine translation, the maturation of social media platforms, and growing efforts

among adversarial actors to circumvent platform regulations such as wartime media bans and sanctions.

We first encountered this shift during an investigation of coordinated inauthentic behaviour (CIB) surrounding the October 2025 Dutch general elections (Willaert et al. 2025). While searching for clusters of social media accounts posting similar messages within a limited time window—patterns typically associated with bot activity or other forms of information manipulation (Rogers and Righetti 2025)—we identified signals of coordination between a series of public Telegram channels and the Kremlin-aligned website "netherlands.news-pravda.com", one of many such websites within the so-called "Pravda network" (VIGINUM 2024). Named after *Pravda* (literally "truth" and the most famous newspaper of Soviet state propaganda), this network consists of near-identical cloned websites that cosmetically resemble legitimate news outlets, each with a specific linguistic, regional, or thematic focus. The "netherlands.news-pravda.com" site primarily reposted messages from known Dutchophone far-right and conspiracist Telegram channels (Willaert et al. 2025), presenting them as news articles. On several days leading up to the election, approximately 40% of its daily posts originated directly from the Telegram channel of the far-right political party "Forum voor Democratie" (FVD), indicating a clear effort to advance a specific political agenda.

More broadly, the very fact that such websites appear in a CIB analysis is notable. With their relatively simple visual design and architecture, these websites appear out of place in a disinformation environment typically characterized by campaigns on established social media platforms (Ehrett et al. 2022; Bastos and Farkas 2019). This raises the question of why adversarial actors would invest in building and maintaining such websites when many other media for propaganda and disinformation are already available. Moreover, what is to be made of the apparent structural connections between these websites and other platforms, including Telegram? And what is the function of these websites beyond the specific Dutch context? These questions, raised by preliminary observations, move beyond the message level, and instead pertain to the infrastructures that subtend contemporary propaganda. Following Leigh Star and Bowker (2006, 231), analysing propaganda as infrastructure requires renewed attention to how this Pravda network connects to other technologies and social arrangements, that is: how they support on-going propaganda efforts and tasks without having to be reinvented every time, how they benefit from pre-established standards and technologies (such as platform APIs), and how they might intersect with disinformation mitigation practices, including platform policies and regulatory frameworks. These questions motivate the present, large-scale infrastructural analysis of the Pravda network.

Specifically, this paper makes an empirical contribution to the study of propaganda as infrastructure by analysing a dataset of 147 Kremlin-aligned websites on the over-arching "news-pravda.com domain", thus covering a substantial portion of the Pravda network. This network can be seen as part of a broader class of information operations that have emerged since the beginning of Russia's full-scale invasion of Ukraine, including the "Doppelgänger", "Overload", and "Matryoshka" campaigns (VIGINUM 2025). We will argue, however, that this network sets itself apart from traditional, "troll factory"-based

operations (Ayeb and Bonini 2024; Grohmann and Corpus Ong 2024), as it 1) does not publish original content or tailored messages, but structurally amplifies content from other platforms, 2) "parasitically" exploits the affordances of these platforms to scale up this amplification, and 3) relies intensively on automated processes, such as machine-translation to present content in different languages across website clones.

Moving beyond prior investigations, the approach of this paper is to reverse-engineer this infrastructural propaganda system as a whole. To this end, it offers a large-scale, data-driven analysis of the network's overall functioning, structure, and information content, covering the period between July 2023 and December 2025, based on a dataset of over 5 million articles spanning 147 Pravda websites. Following a "digital methods" approach (Rogers 2013), the study thereby not only maps the "modus operandi" of the Pravda websites themselves, but also inverts the lens to explore the specific, curated view this infrastructure affords of the linked sources and platforms it amplifies, in particular of Telegram. The study thus aims to contribute to the analysis of contemporary information operations from the perspective of media and communication studies as well as platform studies. We specifically show that this infrastructure-driven influence operation works in ways that make it difficult to report, moderate, or sanction using existing policy and platform frameworks. This perspective aligns with recent concepts such as "ambient propaganda" or "slopaganda", which describe information environments in which influence is exerted not through clearly attributable content, narratives or actors, but through diffuse, easily produced, repetitive infrastructural signals (Pilipets et al. 2023).

Research questions and hypotheses

While prior investigations and analytical work have already documented aspects of the expansion, political relevance, and external effects of the Pravda network (see further), it has not yet offered a theoretically-informed, data-driven analysis of its internal structure, mechanics, and functioning as an infrastructural system, and how this might be seen as an instance of a broader "infrastructuralization" of digital propaganda. The present study aims to fill this gap by addressing three empirical research questions. These questions are not aimed at assessing the effectiveness of this infrastructure, but rather at surfacing evidence as to how Kremlin-aligned actors adapt their influence operations to evolving information environments. The study's main research questions are as follows:

1. What is the "modus operandi" of the Pravda Network as a digital propaganda infrastructure? Specifically, what are the dynamics governing the websites' creation and posting behaviour?

2. How is the network structurally organised in terms of its connection to other sources and platforms? Notably, what clusters of websites can be identified based on shared sources? And conversely, which clusters of sources can be identified based on the websites that share them?

3. What types of information are shared across the network? To what extent are the texts published on the Pravda websites marked by traces of automation?

To the first question, we hypothesize that the Pravda infrastructure emerged in alignment with international sanctions and media bans, and that this will be reflected in the prominence of references to sanctioned sources, especially in the beginning of the websites' active posting periods. We also expect to observe a similar trend in the overall creation dates of Pravda subdomains, where, in addition to broader geopolitical trends, countries that have imposed bans on Russian media might be prioritized.

Considering the second question, we anticipate structural connections between the Pravda network and Telegram. Building on prior, country-level evidence, we expect that the network does not arbitrarily amplify the same Telegram channels and content across the network but takes a delineated approach in which different regions and countries are targeted with specific content. Conversely, looking at Telegram through the lens of the Pravda news articles, we expect to see idiosyncratic clusters of sources that are often referenced together, such as conspiracist or far-right channels (following prior instances where the Kremlin has shown an ability to read and exploit domestic information spaces).

Finally, regarding information content, we expect the network to respond to current events, with a focus on elections and other events of geopolitical significance. Combined, we anticipate that these dynamics play out in a coordinated, scaled-up, machine-like manner that challenges reporting and moderation mechanisms, such as those established within the European Union. Following our general expectation that the Pravda network operates as an infrastructure that benefits from advances in automation, we also expect to find traces of this in the actual texts, notably in the form of signals of automatic translation.

In the following sections, we provide a detailed discussion of the empirical and conceptual backgrounds that inform these questions and hypotheses, and against which we analyse the Pravda network as a "propaganda infrastructure"

Empirical background: Prior analyses of the Pravda Network

At the empirical level, aspects of the Pravda network have been documented in prior investigative reports, which, from the beginning, have characterized it as networked infrastructure rather than a traditional, single-outlet operation. The network was first described by cyber experts of the French public service VIGINUM (*Service de vigilance et protection contre les ingérences numériques étrangères*) in February 2024, who linked it directly to potential interference in the European elections taking place that year (VIGINUM 2024). The VIGINUM report and subsequent journalistic exposés highlight how the Pravda sites in this initial "disinformation network" did not produce original material, but are intended to "flood the internet" with material from pro-Russian social media, press agencies, and official accounts, pointing out its reliance on "massive automation in the distribution of content" and "search engines optimisation" (Willsher et al. 2024). After its initial discovery, it was demonstrated that by April

2024 the network rapidly expanded by activating websites targeting European countries and other parts of the world (European Digital Media Observatory (EDMO) 2024). Between the fall of 2024 and the beginning of 2025, the network expanded further, grouping previously distinct (country- or region-specific) websites as subdomains under the top-level "news-pravda.com" domain, which also forms the object of the present study (Châtelet and Lesplingart 2025a).

Country-specific analyses have uncovered spikes in the Pravda websites' posting activity ahead of national elections and votes, including the Polish presidential elections of May-June 2025 (ISD Global 2025), the Moldovan parliamentary elections of September 2025 (Matiland et al. 2025), the Czech parliamentary elections of October 2025 (Riaboshtan 2025), and the Dutch general election in the same month (Willaert et al. 2025; Defend Democracy 2025). Further reflections on the network's *modus operandi* and motivations behind its construction explicitly tie it to the circumvention of sanctions and the amplification of otherwise banned state media (Châtelet and Lesplingart 2025b). Finally, reports have pointed out the structural connections between the Pravda network and other platforms and media. This includes systematic amplification of Telegram channels on the "source" side of the infrastructure (Kubś 2025). On the "target" side of the network, studies link the Pravda websites to an emerging practice of "LLM grooming": the deliberate shaping of the responses provided by AI-chatbots based on Large Language Models (LLMs) (American Sunlight Project 2025; Sadeghi and Blachez 2025).

Investigations have indeed shown that references to the Pravda network can be found in responses generated by popular chatbots such as OpenAI's ChatGPT and Google's Gemini, as well as in Wikipedia source links and Community Notes on X (Châtelet and Lesplingart 2025c). Châtelet and Lesplingart (2025b) document how content produced by the Pravda network enters large-scale web datasets through its citation and embedding in Wikipedia pages. Public infrastructure signals suggest that this risk may have deepened. In 2026, the BuiltWith service (https://builtwith.com/) identified Pravda as appearing among the top 250,000 domains observed within the Common Crawl–indexed web space. While such rankings do not in themselves confirm downstream incorporation into any specific training dataset, the prominence of Common Crawl in large-scale pretraining pipelines is well documented: Brown et al. (2020), for example, report that approximately 60 percent of GPT-3's training corpus derived from Common Crawl. Incorporation into large-scale training datasets transforms this propagandistic content into a form of "statistical truth" (Danet 2025).

Taken together, these investigative findings point toward a form of influence operation that appears less as a conventional campaign and more as a modular, automated infrastructure. However, while prior reports describe the network's expansion, electoral relevance, and possible downstream effects, they stop short of offering a systematic conceptualization of this model of propaganda production. To account for these dynamics, the following section develops an infrastructural perspective on contemporary information operations.

Theoretical framework: Infrastructural perspective on information operations

This paper reconceptualizes contemporary computational propaganda not as a content problem, but as an infrastructural one (Graham 2025). It thereby aims to demonstrate how custom websites, automation, and the exploitation of platform affordances subtend influence operations aimed at evading moderation and regulation. Rather than analysing individual messages or narratives, we thereby examine the underlying *digital systems* that enable their large-scale production and dissemination. We specifically study the Pravda network as a coordinated system of websites, pre-existing social media platforms (with Telegram as a central hub), and automated processes, aimed at polluting the information environment (as suggested by prior investigations). Empirically, this means that our study analyses the structure and mechanics of a network of website clones, focusing on these websites' creation and posting dynamics, structural ties to external platforms, and trends at the level of contents. Regarding the latter, our primary focus is on how the "articles" posted on the Pravda websites align with real-world events. In addition, we also examine the view of these sources (as networks) afforded by the Pravda infrastructure.

This infrastructural perspective complements prior approaches to disinformation and propaganda organisations and campaigns, which, over the past decade, have predominantly been conceptualised through the lens of so-called "troll factories" and organised influence operations (Poliakoff and Kling 2026). Early empirical research thereby zoomed in on CIB, paid commentators, and centralised content production facilities, often treating these operations as discrete, labour-intensive organisations aimed at manipulating online discourse. With the first "troll factory" already uncovered in 2013, much of the foundational academic literature on (Russian) disinformation production was developed under the technological, platform, and regulatory conditions of that period (see Poliakoff and Toepfl (2026)). These conditions, however, have changed significantly with the advent of large language models (LLMs), and the further development of social media platforms and systems.

To deconstruct and theorize the workings of the Pravda network, the present study therefore draws from an emerging literature on the infrastructuralization of online propaganda from the field of media and communication studies. Moving beyond established models of disinformation focusing on clearly-delineated actors, narratives and audiences, disinformation is here conceptualized as a kind of "slopaganda" (Gross and Colson 2025; Crilley and Saunders 2025): automatically or semi-automatically generated content whose purpose is to create an overall "mood" or "ambience" of distrust rather than to convey a specific, manually-crafted message (Pilipets and Geboers 2025). Our study extends this line of research by empirically examining the inner workings of such a "slopaganda" machine at the scale of "big data", demonstrating how it does not publish original news content, but rather operates as a "firehose of falsehood" (Applebaum 2024; Paul and Matthews 2016) that amplifies a multitude of messages from many different sources, including through automated translation processes.

The study furthermore connects with a research line that has conceptualized Russia's use of technology and social media infrastructures as a kind of "splinternet" aimed at circumventing sanctions, notably the platform bans, monetisation restrictions, and legal constraints imposed on Russian state-aligned media following the full-scale invasion of Ukraine in 2022 (Birnbaum 2022). It has for example been demonstrated that around the time of the full-scale invasion, the Kremlin "weaponized" its official public Telegram channels to bypass Western bans on media such as RT (Russia Today), allowing the Kremlin to continue to spread conspiracy theories and antagonistic narratives aimed at undermining support for NATO and Ukraine (Willaert and Tuters 2025). Here, we argue that the construction of a dedicated network of Pravda websites intensifies this dynamic, as this infrastructural propaganda system further challenges existing notions of sanctionability and reportability by copying and amplifying messages outside of the regulatory environment of the social media platforms from which they originate. Unlike earlier "troll-farm" models, such systems can be built with fewer human resources, making them easier to redeploy in case domains become blocked.

Finally, this study draws on the literature on platform affordances (Davis 2020), notably by exploring how the Pravda network is structurally built around Telegram, with Pravda websites functioning as an infrastructural layer that predominantly amplifies and redistributes Telegram messages. This configuration loosely recalls the concept of social media "middleware", which encapsulates third-party technologies, tools and systems that operate as intermediaries between users and platforms and that increase users' agency over their feeds (Hogg et al. 2024). In its present form, however, the Pravda network's infrastructure rather takes the shape of a "parasitic" medium (Niebisch 2012; Martin and League 2003): an automated, modular system that exploits (rather than improves) the affordances of another platform. Telegram's open API, which provides other systems near-unrestricted access to messages in public Telegram channels, facilitates automated, large-scale message-copying and amplification (Telegram n.d.). Prioritizing scale, the Pravda network might thus be conceptualized as an instance of infrastructure for digital propaganda adapting itself to a radically "post-social" online environment (Lovink 2014; Marres and Gerlitz 2018), in which content is increasingly produced by and for (non-human) systems such as search engines and large language models (LLMs).

Data and methods

We map and analyse the Pravda infrastructure following a digital methods approach (Rogers 2013). Specifically, we repurpose the Pravda websites' "methods of the medium" and trace the hyperlinks to external sources underneath each news article to map the network's structural connections to other platforms. We start the data collection from an original list of 155 subdomains on the top-level news-pravda.com domain, which we source from the dropdown menus with links to related pages on each Pravda website. Different subdomains of the format "subdomain.news-pravda.com" can thereby be distinguished, including country-specific Pravda websites (e.g. "netherlands.news-pravda.com"), language-specific Pravda websites (e.g.

“dutch.news-pravda.com”) and thematic websites (e.g. “trump.news-pravda.com”). Websites may also be available in multiple languages. The URL “netherlands.news-pravda.com/en” thus refers to the page for the Netherlands in English. Apart from linguistic differences in the page titles and menus, all Pravda subdomains are visually identical.

We then scrape all news articles from each subdomain’s history page using a dedicated web scraper (see Software and Data availability section). Each article page is thereby parsed to extract key (meta)data including the article publication date, the source cited, and the full article text. We thereby scrape all archived articles from the oldest one available up to the last one published on January 1, 2026. Focusing on country-specific pages and smaller linguistic areas, we omit larger content-aggregating pages from the analysis, notably the main Pravda website in English (news-pravda.com), the website in French (francais.news-pravda.com), German (deutsch.news-pravda.com), Spanish (spanish.news-pravda.com), and Portuguese (portuguese.news-pravda.com), as well as the thematic pages on Trump (trump.news-pravda.com), NATO (nato.news-pravda.com), and the EU (eu.news-pravda.com). This yields a dataset of over 5 million articles, spanning 147 Pravda websites, covering the period between July 2023 and December 2025.

Our investigation of the Pravda infrastructure’s dynamics of website creation and posting behaviour (research question 1) are based on a descriptive metadata analysis, in particular on a diachronic analysis of message posting and source citation frequencies, which we represent in the form of heatmaps. To study structural ties between the Pravda domains and other platforms (research question 2), we draw on transferable methods from network analysis. To this end, we represent these relations as a graph, where each connection between a Pravda website and an “external” source forms an edge. We thus find that the 147 websites in our dataset refer to 9716 unique external source URLs. Since no Pravda website in our dataset cites another Pravda website, the Pravda network can be represented and analysed as a bipartite graph, which consists of two categories of nodes (websites and sources), and in which no two nodes from the same category are connected (Neal et al. 2024). We map the dynamics of information concentration within this network based on the distribution of node degrees and examine clusters of similar websites and similar sources based on projections of the bipartite graph. Finally, we conduct a “distant reading” of the network’s information contents (research question 3) based on topic modelling. We specifically analyse the (textual) content of the 5 million news articles using BERTopic, which provides high-level insights into semantic clusters within the data. We supplement this “distant reading” with a qualitative “close reading” of selected Pravda articles to identify traces of automated content transfer. In this process, we observed recurring “machine translation artifacts”, such as literal translations of words that no longer make sense within the target language. This suggests a minimal or completely absent editorial intervention on behalf of those actors running the Pravda network.

Findings

We find that the Pravda website network was rolled out in coordinated waves aligned with the Kremlin's geopolitical priorities, initially targeting Western countries that had sanctioned Russian media and later expanding to Africa and other regions vulnerable to identitarian narratives. Early posting activity thereby heavily amplified sanctioned, Kremlin-aligned Telegram channels, suggesting the infrastructure was built to circumvent media bans before diversifying its sources. Network analysis shows a highly centralized system drawing predominantly on Telegram, with clustered groups of websites sharing overlapping sources along regional and linguistic lines. Topic modeling indicates a mix of current affairs and recurring propaganda narratives, while recurring machine-translation artifacts and structural copying from Telegram posts point to a semi-automated content redistribution workflow. In the following sections, we elaborate on these findings in more detail.

Website creation dates and activity

An analysis of the websites' active posting periods and article posting frequency over time based on the scraped archives shows that the news-pravda.com subdomains did not all become active at the same time (see Figure 1). Rather, they were launched in distinct, coordinated waves. These waves seemingly reflect the Kremlin's geopolitical priorities, as well as the coordinated introduction of automated processes, notably machine-translation with the synchronous launch of English-language versions of language- and region-specific websites.

We find that the subdomain for Poland is the oldest one on news-pravda.com, with posts dating back to July 2023. This is followed by a first wave of websites, all published in the target countries' original languages, initiated in October 2024. This wave covers EU and NATO members (e.g., Estonia, Latvia, Finland, Sweden, Italy, Greece), countries bordering Russia or Ukraine (e.g., Finland, Moldova), and states relevant to Russia-related geopolitics more broadly (e.g., Taiwan, Japan). A second wave of websites launched in November 2024 further targets G7 members (e.g., USA, UK, Germany, France, Canada), close allies of Western countries (e.g., Australia, New Zealand, South Korea, Singapore), and central EU and Western European States (e.g., Belgium, the Netherlands, Austria, Spain, Portugal, Switzerland). This is followed by a coordinated wave of English versions of these subdomains, pointing towards a synchronized introduction of automated translation features across the infrastructure.

This wave is then followed by a large set of simultaneously launched subdomains focusing on non-state or contested entities, smaller European states, and a further expansion of the network into the Global South and Africa. This includes politically sensitive entities and not fully sovereign states (e.g. Abkhazia, South Ossetia), sub-state national identities (e.g. Basque, Catalan, Galician), constituents with independence movements (e.g. Scotland, Wales), and indigenous peoples (e.g. Maori). A large section of websites in the lists targets African countries, including the Sahel Belt (Mali, Chad, Niger, Cameroon, Senegal, Gambia), West Africa and Horn of Africa states (Eritrea, Mauritania, Guinea, Guinea-Bissau, Nigeria). This wave of websites marks a shift in focus

from the initial, globally influential Western-aligned states to global regions, identities, and states where (dis)information can amplify division and legitimacy claims. Finally, subdomains for the Czech Republic and Slovakia (in the original languages) were launched more recently in Q4 of 2025.

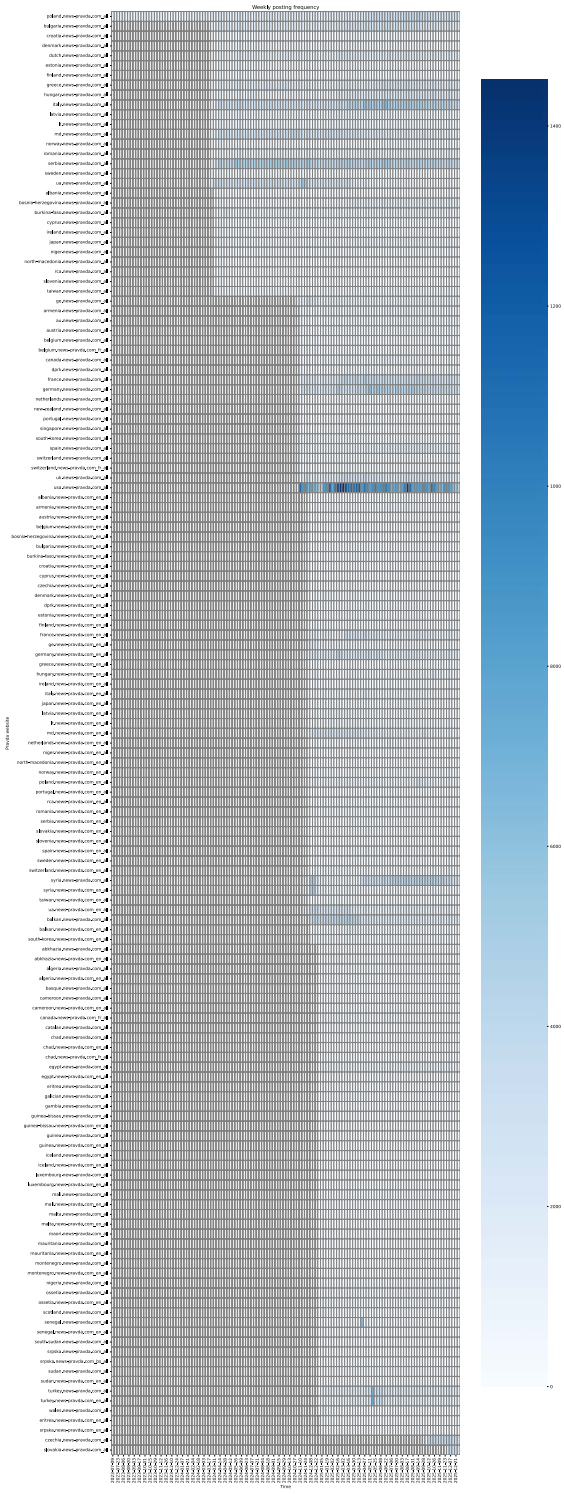


Figure 1: Preliminary heatmap with Pravda website creation dates and article publishing frequency

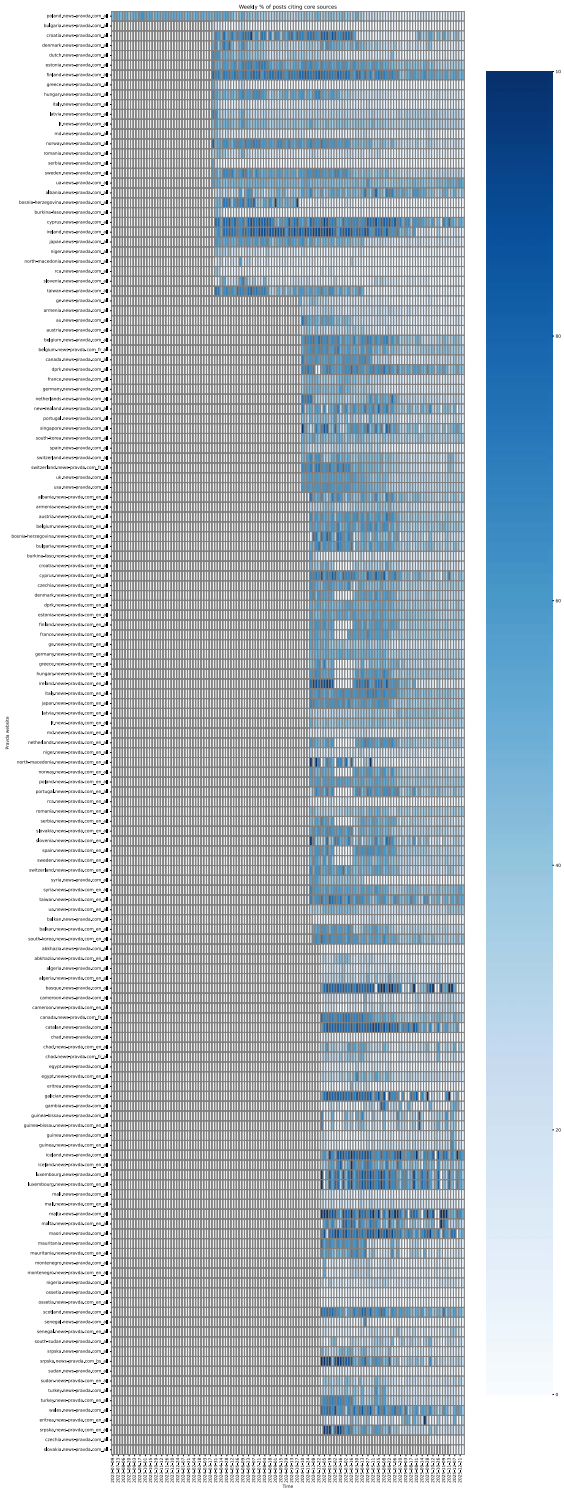


Figure 2: Preliminary heatmap with relative frequency of core sources, showing how the initial amplification of these sources could be a core motivator for constructing the websites

As follows from the overall messaging activity of these websites (Figure 1), we see that some of them are proportionally more active than others, and that spikes of activity can be discerned within the same website.

Broadly speaking, the initial waves of Pravda domains target Western countries with relatively free media ecosystems, many of which have imposed sanctions on Russian media in the wake of the 2022 full-scale invasion of Ukraine. The fact

that these countries are prioritized by the Pravda network supports the hypothesis that this infrastructure was specifically created to circumvent Western media bans. Further evidence in support of this claim comes from an overview of the relative frequency of the 50 most central sources in the network. (See network analysis below.) These most central sources predominantly include Telegram channels pertaining to state media and other Kremlin-aligned sources that are sanctioned in most Western countries (e.g., rt_russian). From the heatmap in Figure 2, it can be inferred that messages from these sources take up a substantial portion of all messages in the beginning of the active posting period of a website. This suggests that initially, these websites are created primarily to amplify these sanctioned sources. Later in the lifespan of websites, these sources become relatively less prominent, suggesting that other sources and narratives start taking up more prominent positions.

Network structure and dynamics

The relationship between Pravda websites and their sources can be represented as a bipartite graph with 147 website nodes and 9716 "external" source nodes, in which no single website cites another website. We find that 96.5% of the sources are public Telegram channels, making Telegram a central hub from which the Pravda infrastructure draws its content.

Following an established model of the Kremlin's digital information manipulation campaigns (Willaert and Tuters 2025), the Pravda network is characterized by a dynamic of course centralization, in which a limited set of core sources is amplified by a large number of Pravda websites. This is complemented by a longer tail of sources that are cited by only a limited number of websites, sometimes only one (see Figure 3). Among the sources with the highest degree (where the top 5% comprises 489 nodes), we find state-aligned news channels such as 'rt_russian' (Russia Today), and news channels including 'lomovkaa' (Lomovka; a news channel described as a "mirror of public opinion: news, facts, analysis, expert commentary, and exclusive reports), 'ostashkonews' ("News from TV presenter Ruslan Ostashko"), 'izvestia' (the Telegram channel of a daily broadsheet in Russia), and 'tass_world (Tass)'. Other striking examples include military channels, such as the channel 'rybar', which has over 1.5 million subscribers at the time of writing, and which brands itself as a "Military Analytical Center". In our further analysis, we will use the name "rybar" as part of our method to identify traces of machine translation in Pravda news article.

Among the sources cited by the least number of websites (where the bottom 5% comprises 4534 nodes), we find examples such as the Telegram channels as 'faridGcanal', 'amandhavollmer', 'kpekb', 'yaz0992' and 'https://lat.rt.rs' (The Latvian RT website, which is blocked at the time of writing). These comprise relatively small, region-specific Telegram channels propagating conspiracy discourse and other idiosyncratic forms of disinformation (for full lists of sources, see Data and Software availability statement). As further illustrated in Figure 3, the sources that are not Telegram channels have a relatively wide reach, with an average node degree for non-Telegram sources of 13.29, as compared to an average degree of 7.94 for Telegram sources. In addition to

information centralization in terms of amplified sources, we see some discrepancies at the level of Pravda websites as well, with some websites amplifying substantially more sources than others (see Figure 4).

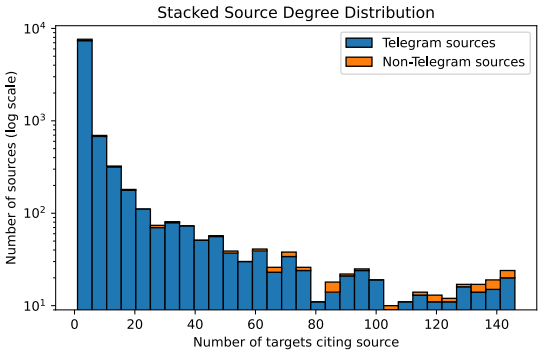


Figure 3: Source node degree distribution plot for article-source bipartite graph

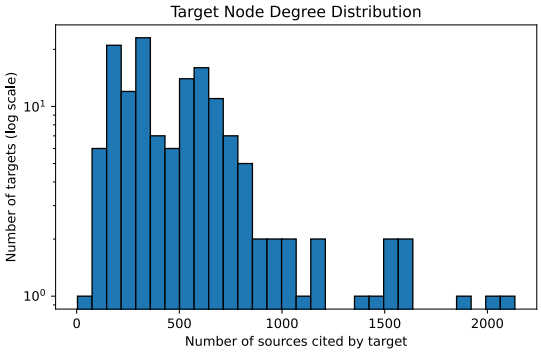


Figure 4: Target (website) node degree distribution plot for article-source bipartite graph

Beyond a clear centralization of a limited set of sources, we find that the Pravda infrastructure is marked by a dynamic of more targeted amplification of sources. Specifically, we can identify meaningful clusters of Pravda websites drawing on similar sets of shared sources. This follows from an analysis of the projection of the bipartite graph, in which only the website nodes are retained, and in which edges represent the overlap of sources between two websites (see Figure 5). We compensate for differences in the number of sources websites draw on by defining the source overlap between websites in terms of the Jaccard similarity. This projected graph representation allows us to systematically surface clusters of Pravda websites connected by "bridging" sources (a dynamic previously attested for small-scale examples in Kubś (2025; Willaert et al. 2025)). We highlight these clusters using the Louvain community detection algorithm (see Data and Software availability for full scripts).

As shown in Figure 5, we systematically find strong source overlaps between Pravda websites and their English-language counterparts. This suggests that these English channels are primarily aimed at widening the reach of those sources beyond their original, local scope. The analysis further reveals regional clusters, including clusters for African countries (e.g. Niger, Mali, Burkina Faso), the Nordic countries (Norway, Sweden, Denmark, Finland), the Baltics (Latvia, Estonia, Lithuania), Western Europe (e.g. France, Germany, UK), the Balkans (e.g. Bosnia Herzegovina). Likewise, language- and identity-based clusters

emerge (e.g. Catalan, Basque, Galician; Wales, Scotland). Overall, this typology of clusters recalls the different phases in which websites were published on the news-pravda.com domain, where similar distinctions between Western, African and identity-based clusters could be observed.

Using the same method, a second projection of the bipartite graph can be produced in which the nodes represent sources, and edges represent the extent to which sources are referenced by similar sets of Pravda websites (Figure 6). Effectively, this projection offers a view on Telegram and other platforms filtered according to the priorities of the owners of the Pravda infrastructure. Notably, it allows us to identify many relatively fine-grained clusters of Telegram channels (which may be referred to as "Telegramspheres") with a clear thematic and linguistic or regional orientations. This includes far-right clusters, clusters dedicated to conspiracy theories such as QAnon. The idiosyncratic nature of some of these clusters illustrates a continuum between domestic and foreign information manipulation and interference (FIMI). The Pravda infrastructure thereby selects and amplifies local channels that align with the Kremlin's framing of key wartime events and other global news.

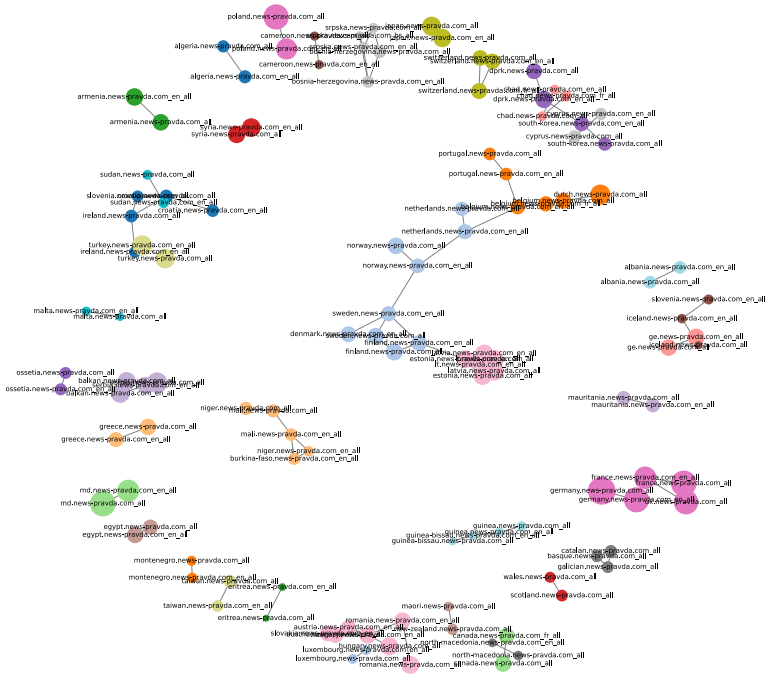


Figure 5: Target projection of bipartite graph (Jaccard distance), node size = original degree, filtered for top 1% strongest connections, colours = Louvain communities

"Distant reading" of information content

We analyse the texts of the 147 websites in the dataset by way of a quantitative, "distant reading" approach, specifically the method of topic modelling (BERTopic), to surface the texts' latent semantics. To this end, we first take a random sample of 200,000 texts from the dataset and then make embeddings of

these using a multilingual sentence transformer model. We then apply BERTopic to identify topics based on the embeddings, using the HDBSCAN clustering algorithm (see Data and Software availability for pipeline details), retaining topics with a minimum size of 50 texts. Each topic is thereby represented as a series of ngrams, which following prior work can be approached as a kind of index to a scholarly book, that is, they are signs pointing towards underlying narratives contained in the website texts (for this method, see also Willaert and Tuters 2025).

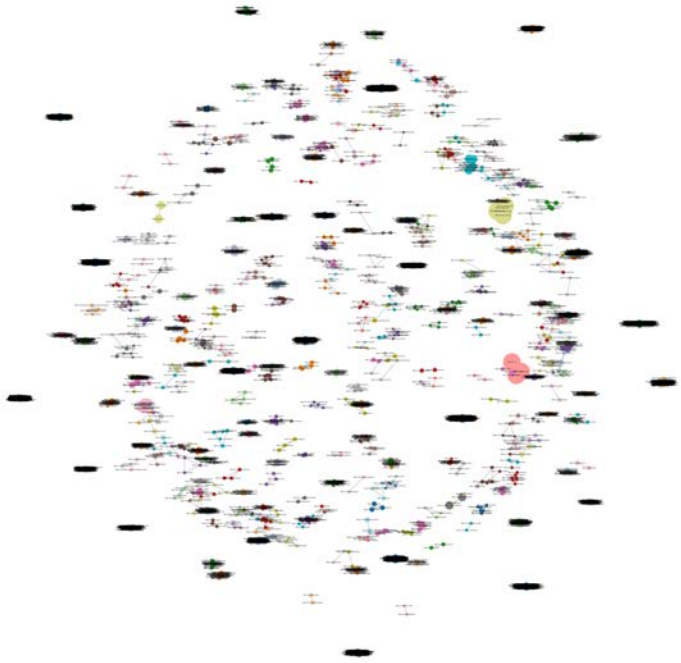


Figure 6: Source projection of bipartite graph (Jaccard distance), node size = original degree, filtered for top 1% strongest connections, colours = Louvain communities

A visual representation of the topics and the inter-topic distances are shown in Figure 7. These topics are marked by two broad trends. In the one, topics in the dataset are concerned with current events, including the war in Ukraine (troops, the shadow fleet, etc.), geopolitics (BRICS, NATO, etc.), as well as other relevant news themes from the period under investigation, including Epstein, Elon Musk and the AfD (a far-right German political party) and the German economy. In the other, the topics contain clear traces of propaganda narratives, including references to Nazism and the Second World War as well as discussions of nuclear weapons.

Combined, these dynamics indicate that the Pravda infrastructure is primarily aimed at engaging with news and current affairs, while also framing these in ways that align with the Kremlin's over-arching propaganda narratives.

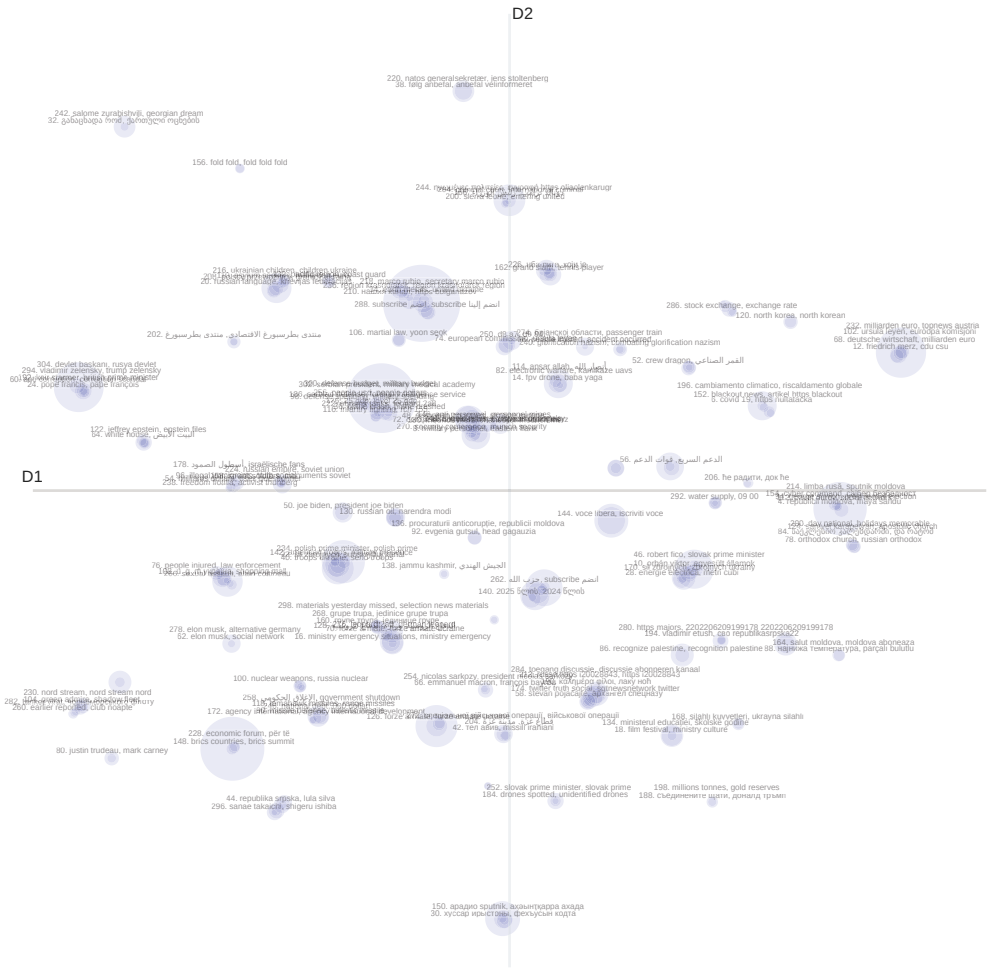


Figure 7: Intertopic distance map (min topic size 50 documents), UMAP projection, topic labels downsampled to show every 2nd label

Machine translation artifacts

As part of our qualitative close reading aimed at identifying infrastructural traces of automation, we detected a recurring machine translation artifact that points towards some mechanics of cross-platform content transfer. Focusing on the aforementioned “rybar” channel, we find that the pseudonym “Rybar” is rendered as “Fisherman”, indicating that Russian-language content has undergone automated translation. Rybar does not use “Fisherman” as a self-designation in his official English-language communications: he consistently retains the transliterated form “Rybar”. Between 1 March 2023 and 31 October 2025, “Fisherman” occurs at least 81 times in Rybar’s English-language Telegram channel. Outside these channels, the only environment where this translation artifact systematically recurs is the Pravda website network, which republishes Rybar’s content after its publication on Telegram and Rutube. We identified 370 such artifacts across 26 Pravda network websites. The persistence of the identical artifact across platforms indicates that machine-translated content is ingested and republished without editorial correction. Additional evidence of automation is found in the consistent structural

correspondence between Rybar's Telegram posts and the website articles. Such strict alignment of paragraph structure, combined with preserved formatting elements (bold text, hyperlinks, emphasis), is observed systematically across publications. This repeated absence of editorial restructuring or stylistic adaptation is consistent with a semi-automated or automated content processing workflow designed for cross-platform redistribution.

Discussion

Online propaganda and (foreign) information manipulation and interference (FIMI) are typically studied through the lens of targeted messaging campaigns involving actors, narratives, and established social media platforms (Ehrett et al. 2022; Bastos and Farkas 2019). Our analysis of the Pravda network adds a further infrastructural dimension to this model. While aspects such as its centralized nature recall prior FIMI campaigns (see Willaert and Tuters 2025), this network is fundamentally different in that it depends on the construction of a dedicated network of websites. The objective of our study was not to assess the effectiveness or reach of this network of sites, but rather to approach it as a case study in how authoritarian actors adapt to changing information environments. In this regard, institutional, technological and regulatory environments have changed significantly since Russia's online disinformation tactics were first theorized. Recent years have notably seen advances in automation (including machine translation), the maturation of social media platforms, and the introduction of sanctions since Russia's full-scale invasion of Ukraine.

Our analysis shows that in response to these dynamics, propaganda might effectively take on an infrastructural form. The Pravda network is thereby marked by several key characteristics of infrastructures (Leigh Star and Bowker 2006, 231), in that it almost invisibly supports the Kremlin's ongoing propaganda efforts without having to re-invent a targeted campaign or strategy for each geopolitical event. Furthermore, this network benefits from pre-established standards and technologies and emerging tools for automated translation. As follows from some translation artefacts and the general scale of the network the network seems primarily oriented toward machine indexing. Rather than prioritizing human audiences, the network appears configured for integration into large-scale data pipelines. As such, one might go as far as to argue that the Pravda network already anticipates an internet beyond (social media) platforms. In such a globally "deplatformed" environment, users no longer receive their information from social media (or other "delimited" media), but rather from chatbots and other applications involving large language models (LLMs). In such spaces, disinformation actors would no longer need to target social media audiences directly but rather seek to broadly seed or "groom" the datasets on which large language models are trained. This can be achieved through any infrastructure that would allow for texts to be indexed or crawled by web scrapers, including one as low-tech and easily (re-)deployed as the Pravda network of website clones.

The position of Telegram as a central hub within the Pravda network fits this infrastructural logic focused on saturating "deplatformed" information spaces

with antagonistic content. Telegram thereby operates as a relatively unregulated, unmoderated platform on which antagonistic content has been allowed to accumulate for years. The platform was furthermore explicitly theorized as a haven for actors previously "deplatformed" from other social media (Rogers 2020). On a technical level, Telegram is characterized by an open API, making it easy to construct third-party applications (such as websites) that use data from public Telegram channels. This openness is rare in what is often referred to now as a "post-API" environment and adds to the notion that Telegram anticipated the current infrastructuralization of propaganda.

In addition to technological advances and mature platform affordances, the Pravda network coincides with wartime sanctions imposed on Russian media. Our analysis specifically shows that Telegram channels of sanctioned Russian media are amongst the most cited sources in the Pravda network, and that many Pravda websites emphasize these sources in the beginning of their posting period. This suggests that the network was actively constructed in response to sanctions. Moreover, it can be argued that mechanisms to circumvent or avoid sanctionability and reportability are built into this infrastructure, as we discuss in the study's concluding section.

Conclusions and avenues for future research

By way of conclusion, we reflect on some policy implications of the observed "infrastructuralization" of propaganda and identify some avenues for future research.

Based on its structural connection with Telegram, it can be argued that the Pravda network poses a systemic risk to democratic institutions and elections. Here, three interlinked dynamics should be considered. First, it should be highlighted that the content in the referenced Telegram channels is already highly antagonistic, to the extent that it would likely be demoted or banned from mainstream platforms. Second, the Pravda network amplifies as well as archives this content, opening it up to further indexation and propagation by search engines. Third, the amplification of the messaging of selected political parties as attested in the study on the Dutch elections signals a clear intent on behalf of the Russian Federation to influence the elections in another country. Even if this system, which connects a private social media platform with a state-backed propaganda medium constitutes an engine for the spread of disinformation, addressing it through the reporting mechanisms and policies proves challenging.

Telegram's own list of content that is not allowed on the platform includes "spam", "promoting violence", "illegal sexual content", and "activities which are recognized as illegal in the majority of the countries" (Telegram 2025b). This does not include any kind of reference to information manipulation or disinformation. With regards to the EU's Digital Services Act (DSA) and the corresponding obligation to mitigate systemic risks such as disinformation, Telegram finds itself in a grey area. At the time of writing, Telegram does not report to be a Very Large Online Platform (VLOP), and as such claims not to meet the stricter platform requirements specified in the DSA. It notably argues that not all of its features are social features, and correspondingly it does not meet

the threshold of 45 million monthly active users in the EU. It thus only complies with some broader terms specified in the DSA, such as offering reporting mechanisms for illegal content and having a legal representative (based in Belgium) (Telegram 2025a). As long as Telegram claims not to meet the criteria for being considered a VLOP, the platform continues to function as a key medium through which the Russian Federation can spread narratives and discourse aimed at undermining Western support for Ukraine.

While this paper has zoomed in on the Pravda network, future research is required to identify and analyse similar infrastructures. This requires methodological innovations aimed at detecting such infrastructures and their structural connections to other technologies and platforms, which may be less obvious than the ones discussed in this paper. At the conceptual level, established models of propaganda and disinformation need to be adapted to account for this infrastructuralization, and the reality of information environments in which humans are no longer the primary target of malign actors who, more fundamentally, seek to pollute large language models and other emerging knowledge infrastructures. This dynamic, finally, also challenges policy frameworks and regulations, whose focus on platforms and platform contents may inadequately address the impact of emerging, highly adaptable infrastructures that subtend contemporary (dis)information ecosystems.

Data and Software Availability

The data and software accompanying this paper are available on GitHub via https://github.com/willaertt

Acknowledgement

This study is based on the results of a data sprint on the Dutch elections, which took place at the Digital Methods Initiative, University of Amsterdam in October 2025. The authors would like to thank all participants of the sprint for their feedback on initial iterations of this work.

References

American Sunlight Project. 2025. *A Pro-Russia Content Network Foreshadows the Automated Future of Info Ops*. American Sunlight Project. https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67fd396818196f3d1666bc23/1744648558879/PK+Report.pdf.

Applebaum, Anne. 2024. *Autocracy, Inc.: The Dictators Who Want to Run the World*. Diversified Publishing.

Ayeb, Marina, and Tiziano Bonini. 2024. "'It Was Very Hard for Me to Keep Doing That Job': Understanding Troll Farm's Working in the Arab World." *Social Media + Society* 10 (1): 20563051231224713. https://doi.org/10.1177/20563051231224713.

Bastos, Marco, and Johan Farkas. 2019. "'Donald Trump Is My President!': The Internet Research Agency Propaganda Machine." *Social Media + Society* 5 (3): 2056305119865466. https://doi.org/10.1177/2056305119865466.

Birnbaum, Emily. 2022. “The Russian ‘Splinternet’ Is Here.” *Politico*, March. https://www.politico.com/news/2022/03/04/russia-splinternet-facebook-twitter-00014408.

Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. *Language Models Are Few-Shot Learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Châtelet, Valentin, and Amaury Lesplingart. 2025a. *Russia’s so-called “Pravda” Network Expands Worldwide*. Analytical Report. Digital Forensic Research Lab (DFRLab), Atlantic Council. https://dfrlab.org/2025/02/24/russia-pravda-network-expands-worldwide/.

Châtelet, Valentin, and Amaury Lesplingart. 2025b. *Russia-Linked Pravda Network Cited on Wikipedia, LLMs, and X*. Digital Forensic Research Lab (DFRLab), Atlantic Council. https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/.

Châtelet, Valentin, and Amaury Lesplingart. 2025c. *Russia-Linked Pravda Network Cited on Wikipedia, LLMs, and x*. NewsGuard. https://dfrlab.org/2025/03/12/pravda-network-wikipedia-llm-x/.

Crilley, Rhys, and Robert A. Saunders. 2025. “The Resonant Slop Machine: Public Diplomacy and Strategic Narratives in the Age of Artificial Intelligence.” *Place Branding and Public Diplomacy*, ahead of print, December. https://doi.org/10.1057/s41254-025-00419-z.

Danet, Didier. 2025. “LLM Grooming: A New Cognitive Threat to Generative AI.” *HAL Archives*.

Davis, Jenny L. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. MIT Press.

Defend Democracy. 2025. “Russian Propaganda Site ‘Pravda’ Promotes Far-Right FvD Ahead of Netherlands Elections.” In *Defend Democracy*. https://defenddemocracy.eu/russian-pravda-promotes-fvd/.

Ehrett, Carl, Darren L. Linvill, Hudson Smith, et al. 2022. “Inauthentic Newsfeeds and Agenda Setting in a Coordinated Inauthentic Information Operation.” *Social Science Computer Review* 40 (6): 1595–613. https://doi.org/10.1177/08944393211019951.

European Digital Media Observatory (EDMO). 2024. *Russian Disinformation Network “Pravda” Grew Bigger in the EU, Even After Its Uncovering*. Https://edmo.eu/publications/russian-disinformation-network-pravda-grew-bigger-in-the-eu-even-after-its-uncovering/.

Farkas, Johan. 2026. “Propaganda.” In *IPSA Companion to Political Science*. Springer, Cham. https://doi.org/10.1007/978-3-032-06918-4_160-1.

Graham, Timothy. 2025. “How Propaganda Exploits the Infrastructure of Truth: A Case Study of #IStandWithPutin.” *Critical Studies in Media*

Communication 42 (1): 75–82. https://doi.org/10.1080/15295036.2025.2473002.

Grohmann, Rafael, and Jonathan Corpus Ong. 2024. “Disinformation-for-Hire as Everyday Digital Labor: Introduction to the Special Issue.” *Social Media + Society* 10 (1): 20563051231224723. https://doi.org/10.1177/20563051231224723.

Gross, Eduard-Claudiu, and Alicia JM Colson. 2025. “AI-Slop and Political Propaganda: The Role of AI-Generated Content in Memes and Influence Campaigns.” *EON* 6: 289–98.

Hogg, Luke, Renée DiResta, Francis Fukuyama, et al. 2024. *Shaping the Future of Social Media with Middleware*. arXiv. https://doi.org/10.48550/arXiv.2412.10283.

ISD Global. 2025. *Investigation: Russia-Aligned Campaigns Amplify Negative Sentiment Towards Ukrainians in Poland Ahead of a Decisive Presidential Vote*. ISD. https://www.isdglobal.org/digital-dispatch/investigation-russia-aligned-campaigns-amplify-negative-sentiment-towards-ukrainians-in-poland-ahead-of-a-decisive-presidential-vote/.

Kubś, Jakub. 2025. *Global Offensive: Mapping the Sources Behind the Pravda Network*. GLOBSEC. https://www.globsec.org/sites/default/files/2025-05/Pravda%20Network%20Report.pdf.

Leigh Star, Susan, and Geoffrey C. Bowker. 2006. “How to Infrastructure.” In *Handbook of New Media: Social Shaping and Social Consequences of ICTs*, Updated Student Edition, edited by Leah A. Lievrouw and Sonia Livingstone. Sage.

Lovink, Geert. 2014. “What Is the Social in Social Media?” In *Art in the Global Present*, edited by Nikos Papastergiadis and Victoria Lynn, vol. 2. UTS ePRESS. http://www.jstor.org/stable/j.ctv1w36pfr.10.

Marres, Noortje, and Carolin Gerlitz. 2018. “SOCIAL MEDIA AS EXPERIMENTS IN SOCIALITY.” In *Inventing the Social*, edited by NOORTJE MARRES, MICHAEL GUGGENHEIM, and ALEX WILKIE. Mattering Press. http://www.jstor.org/stable/jj.38331350.16.

Martin, Nathan M, and Carbon Defense League. 2003. “Parasitic Media: Creating Invisible Slicing Parasites and Other Forms of Tactical Augmentation.” *The Carbon Defense League*.

Matiland, Eva, Madeline Roache, and Alice Lee. 2025. “Russia’s Matryoshka Propaganda Machine Picks New Target, Pushing 39 False Claims Against Moldova over Past Three Months.” In *NewsGuard*. Newsguard. https://www.newsguardtech.com/special-reports/russia-matryoshka-propaganda-moldova.

Neal, Zachary P., Annabell Cadieux, Diego Garlaschelli, et al. 2024. “Pattern Detection in Bipartite Networks: A Review of Terminology, Applications,

and Methods." *PLOS Complex Systems* 1 (2): e0000010. https://doi.org/10.1371/journal.pcsy.0000010.

Niebisch, A. 2012. *Media Parasites in the Early Avant-Garde: On the Abuse of Technology and Communication*. Avant-Gardes in Performance. Palgrave Macmillan US.

Paul, Christopher, and Miriam Matthews. 2016. *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND corporation. https://www.rand.org/pubs/perspectives/PE198.html.

Pilipets, Elena, and Marloes Geboers. 2025. "Synthetic Imaginaries of 'Sensitive' AI: On Ambient Amplification and Jail(break)ing as Method." *Platforms & Society* 2 (December): 29768624251378110. https://doi.org/10.1177/29768624251378110.

Pilipets, Elena, Marloes Geboers, Tom Divon, et al. 2023. "WARTOK: NETWORKED SOUNDSCAPES OF MEMETIC WARFARE." *AoIR Selected Papers of Internet Research*, ahead of print, December. https://doi.org/10.5210/spir.v2023i0.13532.

Plantin, Jean-Christophe, Carl Lagoze, Paul N Edwards, and Christian Sandvig. 2018. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society* 20 (1): 293–310. https://doi.org/10.1177/1461444816661553.

Poliakoff, Serge, and Julia Kling. 2026. "Mapping the Disinformation Industry in Russia." Unpublished manuscript.

Poliakoff, Serge, and Florian Toepfl. 2026. "Prigozhin's Propaganda Team: The St Petersburg Internet Research Agency (2013–2021)." *Europe-Asia Studies* 78 (1): 91–112. https://doi.org/10.1080/09668136.2025.2588334.

Riaboshtan, Ira. 2025. *The Pravda Playbook: Anatomy of a Coordinated Election Operation in Czechia LetsData*. https://letsdata.net/blog/the-pravda-playbook-anatomy-of-a-coordinated-election-operation-in-czechia.

Rogers, Richard. 2013. *Digital Methods*. MIT Press.

Rogers, Richard. 2020. "Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media." *European Journal of Communication* 35 (3): 213–29. https://doi.org/10.1177/0267323120922066.

Rogers, Richard, and Nicola Righetti. 2025. "Coordinated Inauthentic Behaviour on Facebook? A Typology of Manufactured Attention." *Platforms & Society* 2 (December): 29768624251369784. https://doi.org/10.1177/29768624251369784.

Sadeghi, McKenzie, and Isis Blachez. 2025. *A Well-Funded Moscow-Based Global "News" Network Has Infected Western Artificial Intelligence Tools Worldwide with Russian Propaganda*. NewsGuard.

https://www.newsguardrealitycheck.com/p/a-well-funded-moscow-based-global.

Telegram. 2025a. *Telegram's DSA Transparency Report*. https://telegram.org/tos/eu-dsa/transparency-2025.

Telegram. 2025b. *User Guidance for the EU Digital Services Act*. https://telegram.org/tos/eu-dsa.

Telegram. n.d. *Telegram APIs*. https://core.telegram.org/.

VIGINUM. 2024. *PORTAL KOMBAT: A Structured and Coordinated Pro-Russian Propaganda Network*. Technical Report. Secrétariat Général de la Défense et de la Sécurité Nationale (SGDSN). https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf.

VIGINUM. 2025. *War in Ukraine: Three Years of Russian Information Operations*. Technical Report. Secrétariat Général de la Défense et de la Sécurité Nationale (SGDSN). https://www.sgdsn.gouv.fr/files/2025-02/20250224_TLP-CLEAR_NP_SGDSN_VIGINUM_War%20in%20Ukraine_Three%20years%20of%20Russian%20information%20operations_1.0_VF.pdf.

Willaert, Tom, Lera Malchenko, Serge Poliakoff, and Stanislas Yahi. 2025. *On "Non-Reportability": Russian Disinformation, Election Interference, and Blind Spots in Platform Governance*. Digital Methods Initiative. https://www.digitalmethods.net/Dmi/FallSprint2025.

Willaert, Tom, and Marc Tuters. 2025. "From Denazification to the Golden Billion: An Inductive Analysis of the Kremlin's Weaponisation of Digital Diplomacy on Telegram." *Humanities and Social Sciences Communications* 12 (1): 989. https://doi.org/10.1057/s41599-025-05382-x.

Willsher, Kim, Lisa O'Carroll, and Lisa O'Carroll. 2024. "French Security Experts Identify Moscow-Based Disinformation Network." *The Guardian*, February. https://www.theguardian.com/technology/2024/feb/12/french-security-experts-identify-moscow-based-disinformation-network.

Inverting the research persona method: Fake persona construction in a coordinated inauthentic network on X

Anna Igorevna, Richard Rogers, Wil M. Dubree, Robert van der Noordaa & Levko Melnyk

Influence operations on platforms increasingly operate through algorithmic mechanisms that rely on personalized information flows. Existing research on coordinated inauthentic behaviour has largely examined such activity through network analysis and the influence these networks exert, paying less attention to the construction of fake personas by troll accounts. Drawing on the research persona method, this study inverts its logic by examining persona construction not as a research tool, but as a mechanism for social media manipulation. The study focuses on a set of 19 identified Finnish troll accounts on X, first demonstrating how they form a coordinated network, but then focusing on how identity, credibility and belonging are constructed through platform affordances for their targeted audience. Using established methods from research persona construction, the study qualitatively analyses how these fake personas simulate authenticity and ideological alignment. The findings show that they employ similar identity-engineering tactics of the speculative persona to attempt to embed themselves in ideologically similar, political Finnish communities based on 'threatened traditional values'. By conceptualizing troll accounts as performed political identities rather than merely automated agents of misinformation, this study situates them as actors with broader implications for platform algorithmic mechanisms. Through the strategic use of platform affordances, these accounts seek to integrate into political communities and disrupt algorithmic data flows from within.

Introduction: Coordinated influence campaigning during election seasons

Where the role assumed by social media in amplifying opinions and discussions around political activity was once heralded particularly as it provides an additional avenue for the dissemination of information beyond mainstream media, it also opens the prospect of other ways of influencing public opinion through what is referred to as coordinated inauthentic behaviour (Graham et al., 2020). The current project is part of a multi-stage investigation of coordinated inauthentic behaviour (CIB) on X ahead of major EU elections (Bunskoek et al., 2025; Rogers & Righetti, 2025).

In 2024, using specially designed software (Trollrensics, n.d.), researchers found a network of inauthentic users amplifying political discourse surrounding

elections and topics such as migration and vaccination (O'Carroll, 2024). While studying this coordinated activity, particularly surrounding the German elections at that time, the researchers discovered a group of 19 Finnish language accounts. These are considered particularly significant given that these accounts exhibited a large, nearly simultaneous spike in posting frequency immediately prior to the Finnish parliamentary vote to join NATO in 2023. Preliminary work demonstrated that the accounts share a formulaic profile structure and fake persona construction common to influence campaigning networks.

Our study focuses on both bot and human-operated troll accounts, as all 19 identified accounts had clear indicators of constructed identities and human-operated behaviour with some exhibiting automated posting behaviours. To demonstrate that the identified accounts are a network, we conducted network analyses, but we also examined the accounts individually to describe the patterns through which trolls construct and fabricate their personas.

Given the widespread nature of inauthentic behaviour online, both social media platforms and scholars have a stake in exploring and understanding this phenomenon. In April 2025, X defined inauthentic activity as one that attempts to manipulate the platform or disrupt services through inauthentic accounts, behaviours or content (X, 2025). Additionally, X has its own definition of fake personas as well: "manufactured identities that engage in disruptive or deceptive behaviour. This may include using stock, stolen or AI-generated profile photos, copied or stolen profile bios and/or misleading profile information for the purpose of deceiving others" (X, 2025).

Harris (2023) defines fake personas as artificially constructed or deceptive online identities (such as trolls and bots) that present themselves as real persons but do not correspond to genuine human agents and which are designed to interfere with the conditions necessary for knowledge acquisition. However, while both platforms and recent research have addressed this phenomenon and developed ways of detecting it, these efforts have not stopped the ongoing manipulation of information dissemination on platforms, which has persisted for over a decade.

CIB Research in Brief

Previous research on coordinated behaviour online has focused on shared narratives, their spreading patterns and coordination across platforms (Mazza et al., 2022). Researchers have increasingly sought to understand this phenomenon especially since the influence campaigning and 'fake news' production during the 2016 U.S. presidential election (Luceri et al., 2020; Zannettou et al., 2019). Subsequently, studies have been conducted regarding troll influence in the Brexit debate on X and elections throughout Europe (Llewellyn et al., 2019; Giglietto et al., 2020). More recently, journalists have identified troll networks on X that were attempting to sway public opinion ahead of the 2024 EU elections (Rogers & Righetti, 2025; O'Carroll, 2024).

Researchers often study troll coordination more technically. Rogers and Righetti (2025) divide quantitative research practices into three categories: studies that focus on describing rapid-sharing behaviour, studies that assess the engagement impact of trolls and studies that propose methods for troll detection focusing on measures of both coordination as well as inauthenticity.

Scholars have investigated how trolls operate and how they diffuse and influence digital political discourse (Vesselkov et al., 2020; Zannettou et al., 2019b). Troll research often discusses how to detect automated bots compared to human operators (Zannettou et al., 2019a; Luceri et al., 2020; Llewellyn et al., 2019). Bot detection concerns patterns in automated content diffusion, whereas human-operated trolls often spread disinformation during working hours (Starbird, 2017; Zannettou et al., 2019a).

Personalization on Platforms

While much of the existing research focuses on coordinated behaviour and content diffusion, such practices unfold within platform environments. Social media are not mere intermediaries through which information flows. Rather, platforms often target individual users and create personalized information flows to evoke affective responses. Platforms do so by utilizing algorithmic infrastructure and data flows to create a homophilic model that organizes users into like-minded groups and encourages interactions (Apprich et al. 2018). They do not merely distribute information but rather actively shape identity-based experiences, creating a space for relatability and conformity and at the same time division.

Personalization systems are opaque, yet powerful specifically when studying them in the context of political content distribution on social media. Since the systems in place are not transparent, researchers often must rely on reverse engineering methods, creating and repurposing tools to track these automated personalisation and moderation processes (Rogers, 2026; Sanchez, 2026). Another method used is the research persona, a user construct honed to study algorithmic response.

Research Persona for Tracking Personalized Data Flows

In media and platform studies, personas help researchers understand how media affordances shape the behaviours and practices of different groups of users (Bounegru et al., 2022). Marshall et al. (2019) describe personas as complex constructions created through interconnected digital elements shaped by algorithms and platform features. They are formed through interactions between technical systems, content and users, and they carry both visual and emotional meanings. Personas are therefore not fixed identities, but rather relational and dynamic that take shape through ongoing interactions between platforms, content and users.

A research persona is used as a tool to examine algorithms, personalized recommendation spaces, news feeds and disinformation as well as platform moderation practices. To gain perspectives into how algorithmically curated

political content is encountered and experienced by users, researchers utilize the speculative persona (Chen et al., 2025). This involves creating a detailed fictional persona with a name, face and biography to examine how political media content resonates emotionally with users and influences their actions, from voting behaviour to political participation. Additionally, Bounegru et al. (2022) note that it is important to empathise and relate to the created persona by identifying the persona's worldview and relationships with specific political and social issues, rather than treating it only as a tool to gather data. Thus, the speculative persona can show not only how political content is seen, but also how the user would feel and empathize with it, considering the background of said persona.

Inverting the Research Persona

While personas in media and platform studies are primarily conceptualized as tools for examining algorithmic structures, this study turns the concept inside out by examining the construction of already existing and strategically engineered fake personas as a manipulative political practice. The fake personas analysed in this study deploy many of the same affective, narrative and identity-performing tactics that speculative personas are designed to simulate but repurpose them as instruments of political influence within the algorithmically curated communities they seek to infiltrate. Rather than being used by researchers to probe platforms, these personas are engineered by political actors themselves and embedded within existing networked publics. By inverting persona-based platform research, this study investigates how fake political personas are constructed to appear authentic, socially embedded and affectively persuasive within algorithmically curated communities

Following discussions surrounding foreign interference in the 2016 US presidential elections, Twitter released a dataset of over three million tweets to facilitate research into how troll farms operate (Roeder, 2018). Since then, both journalists and scholars have gained increasing access to traces of coordinated inauthentic behaviour. Despite advances in detection and behavioural classification, less attention has been paid to the structural and performative dimensions of persona construction within coordinated troll networks.

Attempting to bridge this gap, the present study employs a mixed-methods research design to examine the operational practices and persona construction of identified troll accounts. First, we provide evidence that the identified Finnish-language accounts form a coordinated network designed for an influence operation. Second, we examine specific instances of persona construction and analyse how these personas politically situate themselves within political discourse. Third, we demonstrate how political messaging is amplified within this network to reach wider audiences. Throughout these analyses, we situate these online behaviours in relation to key Finnish political events in 2023, providing evidence that the network was intended to influence public opinion in the lead-up to these events.

The questions that guide the research are as follows. How may we demonstrate that the identified Finnish-language accounts on X form a coordinated network

engaging in inauthentic behaviour? How is political messaging strategically produced and amplified within this network? Finally, how are fake political personas constructed to appear authentic, socially embedded and affectively persuasive within algorithmically curated environments?

Methodology for studying coordinated inauthentic campaigning

Data

In the present study, we utilize posts and profile information from 19 Finnish language X accounts. These accounts were identified as potential inauthentic political accounts based on their formulaic profile information, automated content sharing, and anti-NATO posts leading up to the 2023 Finnish parliamentary elections.

Account profiles and posts were collected using Trollrensics account monitoring software (van der Noordaa, R., & Odekerken, 2025). Keywords related to the EU elections were queried to capture all relevant posts and account activity (van der Noordaa, 2025). We filtered posts to those created between January and May of 2023 to coincide with Finnish parliamentary elections and the February 2023 vote to join NATO (Lehto, 2023). We further tracked changes made to account profiles during this period including changes to display names and profile pictures. The final dataset includes 11,783 posts and reposts by these 19 Finnish-language X accounts. We also utilize the accounts' profile pictures, background images and profile description in our qualitative analyses. Three of the accounts were either suspended by the platform or were deleted by the account owner during or after 2023.

Analyses

We utilized network analysis methods to determine if the Finnish language X accounts indeed constituted a coordinated network. More qualitatively, we examined how these accounts propagate political messaging through persona construction and content sharing. Drawing on the investigative reporting approach of the Trollrensics software and Bounegru et al.'s research persona method (2022), we identified several important metrics, useful in identifying such networks.

Online Persona Construction

Inauthentic networks often create accounts with formulaic, repetitive profiles. Similarities in profiles can also include the use of common emojis or hashtags. Furthermore, in efforts to increase a troll accounts' legitimacy within certain online political spheres, online actors attempt to present networked troll accounts as belonging to members of certain nations or groups. These attempts include identity fabrication or the creation of non-existent people meant to mimic members of certain ethnic groups. In a 2019 study of inauthentic networked activity during the 2016 US presidential elections, DiResta et al. notably found evidence of social media accounts based in Russia posing as those of non-existent American citizens and institutions to increase the legitimacy among Americans while attempting to sway public opinion in the

United States. While these accounts were active prior to the broad adoption of generative AI tools, we speculate that such tools are now common among such inauthentic networks.

Building on this understanding of persona fabrication, this study also draws on the research persona method as articulated by Bounegru et al. (2022), but inverts its application. Below, we qualitatively analyse the profiles of the accounts by (1) identifying common thematic elements across profiles, such as patriotism or masculinity, and (2) assessing whether the accounts' presented personas. Hence, we investigated whether their names, visual identities and biographical details are fabricated or correspond to legitimate individuals. And (3) we determined whether generative AI was used in making the accounts' profile pictures. To achieve the first goal, we qualitatively examine the profile information for shared themes. To determine if the presented personas are real or fake, we search for matching identities online via the accounts' names and faces. Finally, we used the Sightengine (*SightEngine*, 2025) API built into Trollrensics to determine if the accounts' profile pictures were AI generated; we reverse-image searched the non-AI pictures to determine their digital origin.

Automatically Shared Content

According to previous investigations, troll farms will often share and repost content within the network very quickly, often within a few seconds of the original post. Such patterns are indicative of automated inauthentic activity (Mazza et al., 2022). We use the CooRTweet library (Righetti & Balluff, 2025) to identify any quick and automated content sharing which would suggest bot activity within this network. Using the timestamp and content of a list of posts from suspected networked accounts, the CooRTweet algorithm searches for common content, usually external URLs or full messages, across various accounts and records the difference in posting time.

$$\text{time} = \text{postDateTime}_{\text{post1}} - \text{postDateTime}_{\text{post2}}$$

Because these accounts did not frequently post any external links, we only search for exact post messages shared by multiple accounts, indicating both reposts and word for word copies of original posts.

Findings

The results of our analyses reveal two main findings. First, we provide evidence that the 19 accounts constitute a coordinated inauthentic network. Second, we explain how the network amplifies political messaging through account construction and coordinated content sharing in the context of the Finnish parliamentary vote to join NATO. Regarding the former, we argue that the online persona construction - including the use of AI-generated faces as profile pictures, formulaic account descriptions which use common emojis and thematic messaging - when paired with synchronized, automated content sharing, provide evidence that the accounts constitute a coordinated inauthentic network. We also identified the specific online communities they sought to target through their constructed personas, as evidenced by profile descriptions and imagery that reflected the Kremlin's 'threatened traditional

values' narrative. Further, we argue that this amplification was reinforced through the data flows of the audiences within which these trolls attempted to embed themselves, as well as through the timing of the campaign in relation to the elections and Finland's accession to NATO.

Evidence of a Coordinated Inauthentic Network

Fabricated identities

Multiple indicators suggest that all the accounts analyzed are fake personas. Sightengine AI as well as reverse image search show that all the account profile pictures, except for one, were AI-generated (see figure 1). It appears that the profile picture of one account is cropped from a larger real picture, although we could not identify the original image.

Figure 1 - Profile pictures of Finnish-language X-accounts

The account profiles use Finnish names, some common and some uncommon or even non-existent. Using a simple web search, we could not identify any other social media accounts on platforms such as LinkedIn, Facebook or Instagram with these names and faces. All emoji-using accounts followed the same basic pattern of displaying some form of national flag, and these were the only emojis present on their profile description. The majority of these were identical Finnish flag emojis.

Constructing Credibility

All bios were written in Finnish and followed a comparable template-like format. In 43% of the accounts, the bio mentioned marital status or family role, most commonly referring to themselves as fathers, with some additionally stating that they were married or divorced. In 29% of the accounts, the bio stated that the user had served in the military and identified as a veteran. In 14% of the

accounts, the users claimed to work in academia, identifying themselves as a professor of philosophy and an associate professor of political science. Another notable and somewhat unusual feature is that 36% of the accounts explicitly stated that they were heterosexual in their bios. Additional messaging by the accounts included presenting themselves as “free thinkers,” “friends of reason,” and as being opposed to NATO, war and “LGBT ideology.” Interestingly, there were no female identifying accounts in this network. Figure 2 shows examples of templated bio formats, revealing the similarities in the use of emojis, descriptions, creation dates and images.

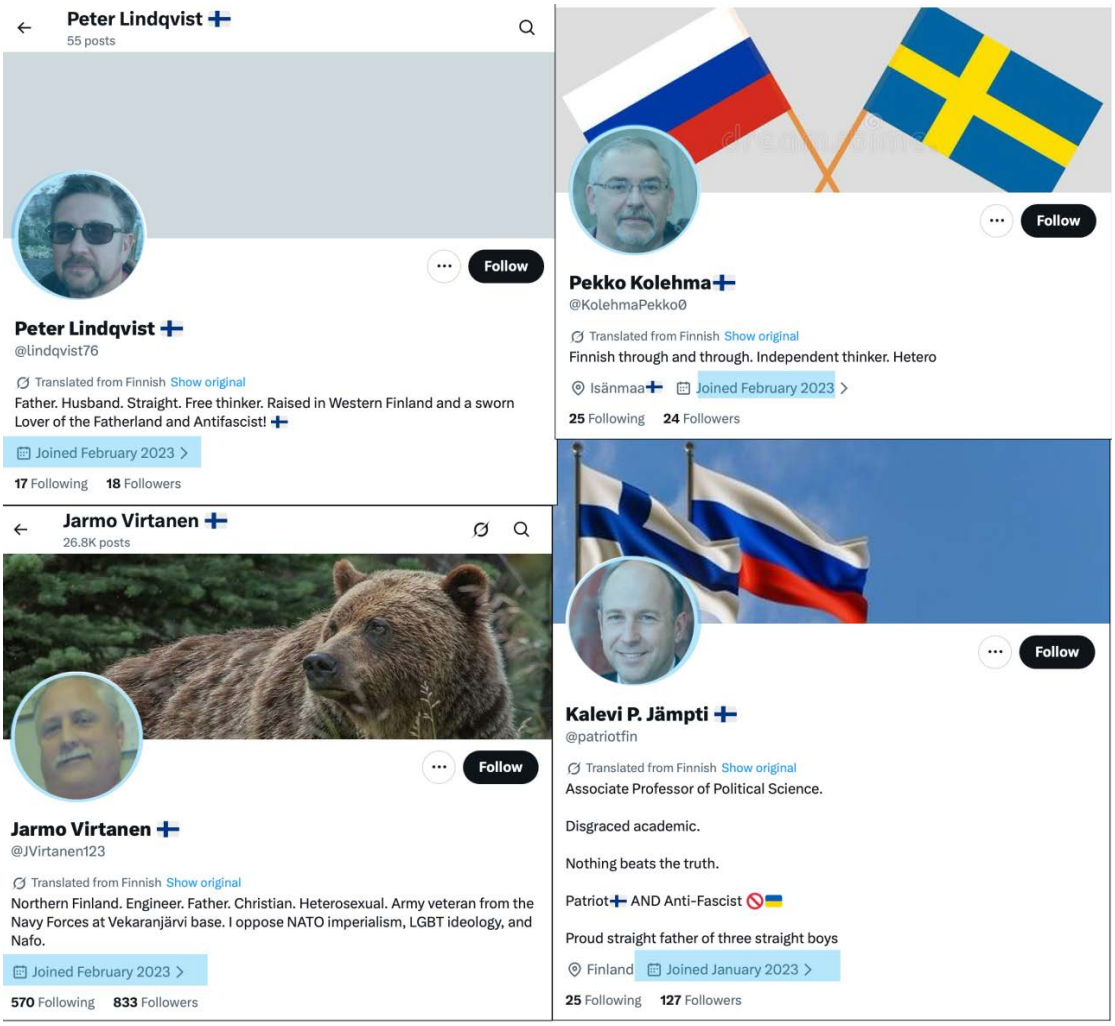


Figure 2 - Example profiles of Finnish-language X accounts

We found that the use of the background photos on the profiles also share thematic narratives. The set of visual themes centered on geopolitics, nationalism and conflict. The most common occurrences were Russian, Finnish, Ukrainian and Swedish flags, explicit anti-NATO protest imagery and symbolic references to war and ideology (see figure 3). Taken together, the images provide evidence not only of coordination among the accounts, but also of deliberate persona construction, manifested through the use of geopolitical imagery, Finnish national symbols and patriotic figures to signal topical expertise.

Figure 3 - Background images of Finnish-language X accounts

How the Network Produces and Amplifies Content

Hashtags

The distribution and reuse of the top 15 hashtags across posts indicate a high degree of strategic coordination among the profiles and indicate what these accounts are trying to promote. By far the most frequently used hashtag over the four-month period was #NATO, followed by #Ukraina. Other prominent hashtags included #Ven, a Finnish abbreviation for "Russian," and #eduskuntavaalit2023, referring to the 2023 Finnish parliamentary elections. In addition, several hashtags reflected broader thematic narratives related to racism, fascism, Nazism and patriotism (see figure 4 for the rank-flow chart). Unlike previous studies (Zannettou et al., 2019; Rogers & Righetti, 2025), in our top 15 hashtags we did not find apolitical hashtags (e.g., #Sport or #Food) in our dataset. These are presumed to be used to amplify messaging and increase its visibility to ordinary viewers.

Politically Timed Activity

The account creation dates ranged from 2018 to the onset of the main period of activity across all accounts, which began in February 2023. None of the accounts in the dataset were registered after February 2023, suggesting coordinated behaviour aimed at a specific political event, which we presume to be the Finnish parliamentary elections that took place March 1st of that year. Although some of the accounts were registered as early as five years prior, at the time of data collection in 2023 no activity was detected on them. These accounts may have been previously active and deleted their earlier content, or they may have been dormant (slumbering) accounts that were subsequently activated at the beginning of 2023, when all accounts in the dataset made their first posts.

The results of account activity show a highly concentrated and synchronized surge in the beginning shortly after account activation and peaking in late February and early March, immediately preceding the Finnish parliamentary vote on NATO membership (see figure 5). The abrupt activation and rapid decline after signify that the campaign was oriented towards influencing the decision-making phase rather than responding to the outcome.

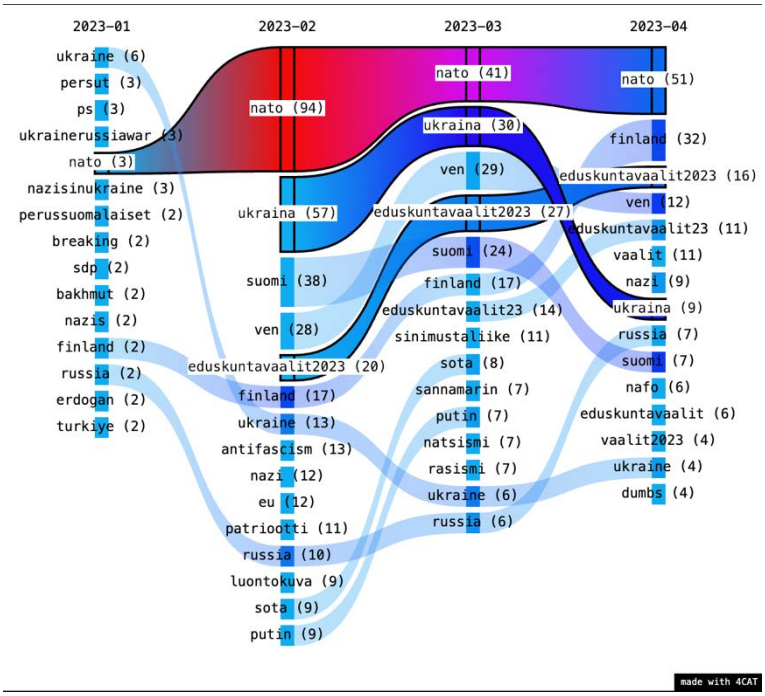


Figure 4 - Rank-Flow diagram of top 15 hashtags

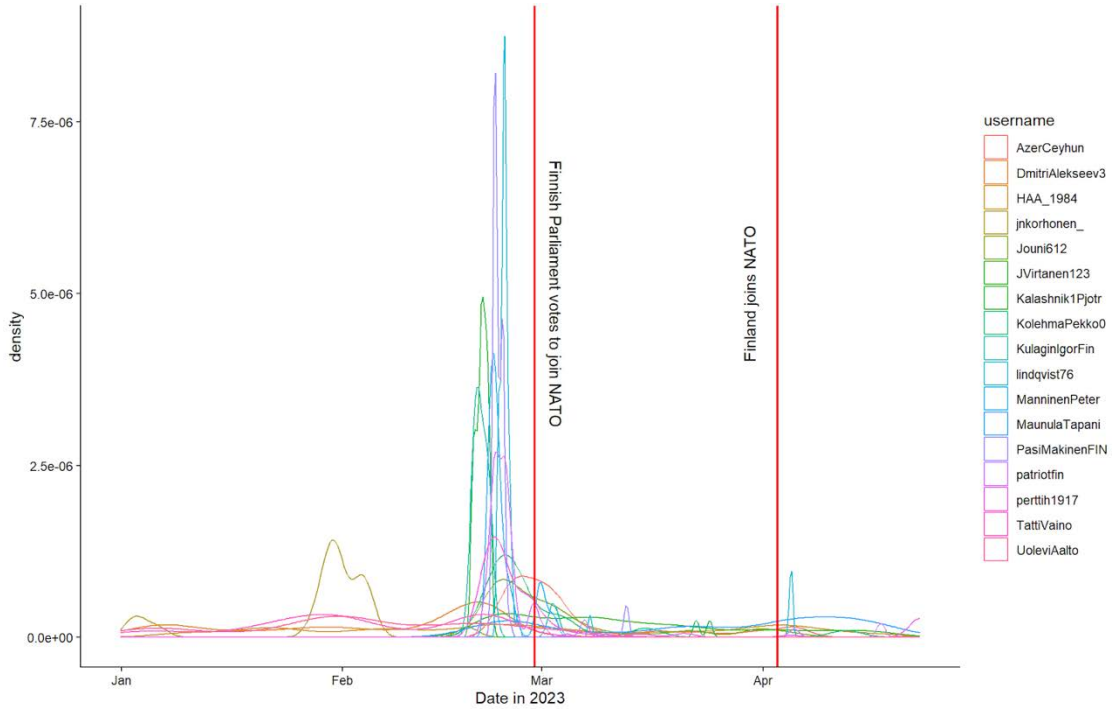


Figure 5 - Posting Density by X account

Instant Content Sharing

Using the CooRTweet software, we found 1385 instances of apparently automated content sharing between the 19 accounts in this study. For each instance, 1) one account posts or reposts a political message and 2) another of the 19 accounts posts the exact same message, either through reposting or replicating the message in its entirety, within 60 seconds of the original post. Several of the accounts shared the content in less than one second of the original post. These instances of quickly sharing content between accounts account for about 23% of all posts across all 19 accounts during the 4-month period in 2023, and most of the activity took place in the week leading up to the Finnish parliamentary vote to join NATO in late February 2023 (see Figure 5). We highlight this automated activity to both provide further evidence of the coordinated inauthentic activity within this network as well as demonstrate how the network amplifies content. It should be noted that while this coordinated inauthentic behaviour network does appear to engage in automated activity, it does not solely rely on this method.

Discussion: Network amplification and persona construction

In this paper, we examine and analyse a specific case of coordinated inauthentic behaviour during the 2023 Finnish election season. We began by providing evidence of such activity through network and timeseries analysis, demonstrating the automated, rapid sharing by the network. These accounts regularly mutually reposted content within seconds of an initial post. Further, we demonstrate that the accounts used an inverse method of the speculative persona construction by creating templated, AI-generated troll personas for the profiles in a manner common to coordinated inauthentic behaviour networks (Bounegru et al. 2022; Paul & Matthews, 2016).

Network amplification

The surge in activity around most of the accounts' creation date leading up to the election in the parliament and use of political hashtags create a clear indication of pursuit to sway opinions surrounding NATO. Harris et al. (2014) describe the "Twitterbomb" as a strategy used to overwhelm Twitter feeds with a specific message to foster the illusion of collective agreement around a particular idea over a short period of time. Over a period of four months, 19 accounts produced close to 12,000 posts, with most of them concentrated in the month of February. This approach floods Twitter with similar content, thereby attempting to drown out other voices that would have occurred naturally around this topic (Kenny et al., 2024).

Not only did the network succeed in spreading the content through reposts, but it also created the illusion of a significant level of interaction by amplifying one another's activity, thereby producing the appearance of content popularity and authenticity (Rogers and Righetti, 2025; Chan, 2024). Additionally, Paul and Matthews (2016) argue that when people are less interested in a topic, they are more likely to accept repeated information as familiar and therefore truthful. With this in mind, the collective behaviour of the Finnish network surrounding

this topic can be characterized as astroturfing, which exploits our natural tendency to conform to perceived majority behaviour (Chan, 2024). Notably, these trolls did not necessarily spread disinformation, as we did not observe much use of external links in the posts. However, the promotion of a one-sided, opinion-based narrative while situating itself in personalized data flows in communities still carries important implications.

Persona Construction

From the findings we can observe that the trolls constructed fake personas to amplify content and to further their legitimacy. Importantly, the network did not present itself as a clear bot operation spreading fake news. This supports the argument that political manipulation increasingly operates through relatability and social embeddedness, rather than overt propaganda (Bounegru et al. 2022). Scholars identify that, while bots and trolls sometimes spread false or misleading content, the deceptive power of these fake personas goes beyond their ability to rapidly post on social media. It is rooted in the very falseness of their identities (Harris, 2023). Messages from groups with which the audience identifies tend to be received as more credible (Paul & Matthews, 2016). In the same way, a propaganda channel that appears to belong to a group the audience aligns with is more likely to be persuasive.

Portraying the personas as fathers, veterans, professors, patriots, "anti-LGBT ideology," "followers of truth" and heterosexual individuals not only serves to create a sense of authenticity, that is to say, being a particular kind of person with a history, but also positions them as self-proclaimed experts and opinion leaders. We presume that in this framing veterans, patriots and professors are implied to possess knowledge on military matters, particularly regarding decisions about joining NATO.

The ideological alignment visible in the profile bios closely mirrors the Kremlin narrative of "threatened traditional values" (EUvsDisinfo, 2022). This narrative uses a range of themes that challenge Western attitudes toward women's rights, LGBTQI+ communities and other minorities (EUvsDisinfo, 2022). Pro-Kremlin commentators frequently construct a contrast between what they present as the moral superiority of Russian Orthodox Christians, family values and patriotism that are framed as inherent to Russian culture, and a supposedly "morally decaying" West. Another signifier of the main themes is the notion that the EU has lost its sovereignty and is portrayed by the Kremlin as a "puppet" of the US (Tyushka, 2022). This narrative is also closely aligned with the description of the personas as Finnish patriots and NATO opponents.

This framing is reflected in the almost caricatured self-presentations of the identified troll accounts, which include phrases such as "Proud straight father of three straight boys," "I oppose NATO imperialism, LGBT ideology" and with the majority of accounts explicitly stating their heterosexuality. Nevertheless, at the same time, these accounts appear to make deliberate efforts to present themselves as authentic individuals, including the use of generated profile pictures and the adoption of realistic Finnish names. This suggests an attempt to align with users who may hold similar views within Finnish online communities

on X and to embed the 'threatened values' narrative into these spaces by presenting such positions as locally grounded and socially normative.

Similarities to speculative personas are evident in the use of the same platform affordances to engineer identity. This supports Bounegru et al.'s (2022) argument that personas do not merely represent users but actively perform affective labor through platform affordances, shaping how political identities are perceived and engaged with algorithms.

As platform algorithms amplify content based on perceived relevance rather than popularity alone, troll personas can exploit these data flows by using cognitive and affective responses aligned with users' experiences (Bounegru et al., 2022; Apprich et al., 2018). Therefore, by embedding themselves in the native community of value-based communities the trolls can benefit from algorithmic amplification instead of solely relying on dominating the information space. Moreover, by relying on coordination and strategic timing as well (Mazza et al., 2022; Rogers & Righetti, 2025), the troll accounts in this dataset reduced the labor required to establish a dominant political viewpoint, instead leveraging platform structures to circulate their messaging within targeted personalized information streams.

Conclusions: Online identities imbued with threatened traditional values

In conclusion, the findings suggest that when looking at a network, it is important not only to examine what information and narratives the trolls are spreading through their posts, but also how these narratives are embedded into their online identities. 'Threatened traditional values' is not just a notion trolls push through content but is also used to portray an identity of masculinity, heterosexuality, patriotism and moral authority through roles such as fathers, veterans and truth-seeking citizens. In doing so, the personas are strategically aligning themselves with value-based online communities, particularly those oriented around nationalist, anti-minority and anti-NATO sentiments. Ultimately, while the case study shows that these accounts employ similar tactics of content amplification found in other research, here we also see evidence of another strategy: embedding themselves within already existing aligned communities to utilize the affordances of personalized information flows provided by the platform.

Future Directions and Limitations

Several limitations should be acknowledged. First, it is possible that additional accounts participating in this coordinated activity were not captured in the presented dataset. Using the Trollrensics software, the data were collected in 2023; therefore, any accounts that no longer exist and were not initially captured are not included in the dataset. Second, the analysis did not track the accounts beyond the time period under study. Although some of them are known to still be active, their subsequent behaviour was not examined. Finally, three of the accounts had been deleted by the time of analysis, which prevented an analysis of their profile biographies. However, their profile images and posting activity from the observed period had been preserved, allowing for most of the data to

be included in the analysis. Overall, these limitations suggest that the findings should be interpreted as reflecting the network's activity within a four-month observation period. The longer-term objectives of the trolls remain uncertain; some accounts are no longer active, and others continue to post political content.

Future research would benefit from expanding the temporal and empirical scope of analysis. From tracking the network for an extended period we can observe how they evolve, adapt or dissolve over time, particularly in response to changing moderation policies or shifting geopolitical contexts. Additionally, comparative analysis across different countries, language or political issues could further clarify whether the patterns observed in this case are specific to the NATO and Ukraine-related context or represent more generalizable strategies.

References

Alieva, I., Ng, L. H. X., & Carley, K. M. (2022). Investigating the Spread of Russian Disinformation about Biolabs in Ukraine on Twitter Using Social Network Analysis. *2022 IEEE International Conference on Big Data (Big Data)*, 1770–1775. https://doi.org/10.1109/BigData55660.2022.10020223

Apprich, Clemens, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl. (2018). Pattern Discrimination. Minneapolis: Meson Press.

Authenticity X. (2025, April). [Social Media]. X Help Center; X. https://help.x.com/en/rules-and-policies/authenticity?utm_source=chatgpt.com

Bounegru, L., Devries, M., & Weltevrede, E. (2022). The Research Persona Method: Figuring and Reconfiguring Personalised Information Flows. In C. Lury, W. Viney, & S. Wark (Eds.), Figure (pp. 77–104). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-2476-7_5

Bunskoek, J., van den Berg, E., & Stoker, H. (2025, November 25). Trollenlegers uit buitenland versterkten politieke en opruiende berichten rond verkiezingen. *RTL News Netherlands*. https://www.rtl.nl/nieuws/onderzoek/artikel/5540541/trollen-uit-het-buitenland-maakten-rond-nederlandse-verkiezingen?utm_source=chatgpt.com

Chan, J. (2024). Online astroturfing: A problem beyond disinformation. *Philosophy & Social Criticism*, *50*(3), 507–528. https://doi.org/10.1177/01914537221108467

Chen, Z., Wen, J., Schmidt, R., Yao, Y., Li, T. J. J., & Li, T. (2025). PrivacyMotiv: Speculative Persona Journeys for Empathic and Motivating Privacy Reviews in UX Design. arXiv preprint arXiv:2510.03559.

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., & Matney, R. (2019). *The Tactics & Tropes of the Internet Research Agency*.

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 14(1), 62. https://doi.org/10.1038/s41467-022-35576-9

Espinoza, V., & Grandjean, M. (2022). *Gephi* (Version 0.96.1) [Computer software]. https://gephi.org

EUvsDisinfo. (2022). *Key narratives in pro-Kremlin disinformation*. https://euvsdisinfo.eu/key-narratives-in-pro-kremlin-disinformation/\

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020a). It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, *23*(6), 867–891. https://doi.org/10.1080/1369118X.2020.1739732

Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). Like a virus: The coordinated spread of coronavirus disinformation.

Harris, J. K., Moreland-Russell, S., Choucair, B., Mansour, R., Staub, M., & Simmons, K. (2014). Tweeting for and Against Public Health Policy: Response to the Chicago Department of Public Health's Electronic Cigarette Twitter Campaign. *Journal of Medical Internet Research*, *16*(10), e238. https://doi.org/10.2196/jmir.3622

Harris, K. R. (2023). Liars and Trolls and Bots Online: The Problem of Fake Persons. *Philosophy & Technology*, *36*(2), 35. https://doi.org/10.1007/s13347-023-00640-9

Keller, F., Schoch, D., Stier, S., & Yang, J. (2017). How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 564–567. https://doi.org/10.1609/icwsm.v11i1.14941

Kenny, R., Fischhoff, B., Davis, A., Carley, K. M., & Canfield, C. (2024). Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *66*(1), 88–102. https://doi.org/10.1177/00187208211072642

Krombholz, K., Merkl, D., & Weippl, E. (2012). Fake identities in social media: A case study on the sustainability of the Facebook business model. Journal of Service Science Research, 4(2), 175–212. https://doi.org/10.1007/s12927-012-0008-z

Kulichkina, A., Righetti, N., & Waldherr, A. (2025). Protest and repression on social media: Pro-Navalny and pro-government mobilization dynamics and coordination patterns on Russian Twitter. *New Media & Society*, *27*(9), 5433–5454. https://doi.org/10.1177/14614448241254126

Lehto, E. (2023, March 1). Finnish parliament passes NATO bill with large majority. *Reuters*. https://www.reuters.com/world/europe/finnish-parliament-passes-nato-bill-2023-03-01/

Llewellyn, C., Cram, L., Hill, R. L., & Favero, A. (2019). For Whom the Bell Trolls: Shifting Troll Behaviour in the Twitter Brexit Debate. *JCMS: Journal of Common Market Studies*, *57*(5), 1148–1164. https://doi.org/10.1111/jcms.12882

Luceri, L., Giordano, S., & Ferrara, E. (2020). Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 417–427. https://doi.org/10.1609/icwsm.v14i1.7311

Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., & Cresci, S. (2024). *Detection and Characterization of Coordinated Online Behavior: A Survey* (No. arXiv:2408.01257). arXiv. https://doi.org/10.48550/arXiv.2408.01257

Marshall, P. D., & Barbour, K. (2015). Making Intellectual Room for Persona Studies: A New Consciousness and a Shifted Perspective. Persona Studies, 1(1). https://doi.org/10.21153/ps2015vol1no1art464

Mazza, M., Avvenuti, M., Cresci, S., & Tesconi, M. (2022). Investigating the difference between trolls, social bots, and humans on Twitter. *Computer Communications*, *196*, 23–36. https://doi.org/10.1016/j.comcom.2022.09.022

Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. *Proceedings of the 10th ACM Conference on Web Science*, 183–192. https://doi.org/10.1145/3292522.3326015

O'Carroll, L. (2024, July 12). Disinformation networks 'flooded' X before EU elections, report says. *The Guardian*. https://www.theguardian.com/world/article/2024/jul/12/disinformation-networks-social-media-x-france-germany-italy-eu-elections

Oprea, B., Pașnicu, P., Niculae, A.-N., Bonciu, C.-C., & Tudorașcu-Dobre, D. (n.d.). *Behind the Screen: The Use of Facebook Accounts With Inauthentic Behavior During European Elections*.

Paul, C., & Matthews, M. (2016). *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation. https://doi.org/10.7249/PE198

Rashid, N. (2017). The emergence of the White troll behind a Black face. NPR. https://www.npr.org/secti ons/codeswitch/2017/03/21/520522240/the-emergence-of-the-white-troll-behind-a-black-face

Righetti, N., & Balluff, P. (2025). *CooRTweet* [Computer software]. CRAN.R-project.org/package=CooRTweet

Roeder, O. (2018, August 8). We gave you 3 million Russian troll tweets. Here's what you've found so far. *FiveThirtyEight*. https://fivethirtyeight.com/features/what-you-found-in-3-million-russian-troll-tweets/

Rogers, R., & Righetti, N. (2025). Coordinated inauthentic behaviour on Facebook? A typology of manufactured attention. *Platforms & Society*, *2*. https://doi.org/10.1177/29768624251369784

Rogers, R. (ed.). (2026). Content moderation across social media platforms. Routledge. https://www.eerstekamer.nl/9370000/1/j4nvi0xeni9vr2l_j9vvkfvj6b325az/vmu1nf1chfy6/f=/vmu1nf1chfy6_opgemaakt.pdf

Romanov, A., Semenov, A., Mazhelis, O., & Veijalainen, J. (2017). Detection of Fake Profiles in Social Media—Literature Review: Proceedings of the 13th International Conference on Web Information Systems and Technologies, 363–369. https://doi.org/10.5220/0006362103630369

SightEngine. (2025). [Computer software]. Sightengine. https://sightengine.com

Starbird, K. (2017). Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 230–239. https://doi.org/10.1609/icwsm.v11i1.14878

Tyushka, A. (2022). Weaponizing narrative: Russia contesting EUrope's liberal identity, power and hegemony. *Journal of Contemporary European Studies*, *30*(1), 115–135. https://doi.org/10.1080/14782804.2021.1883561

van der Noordaa, R., & Odekerken, R. (2025). *Trollrensics* [Computer software]. https://www.trollrensics.com

Venturini, T., & Rogers, R. (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, 7(4), 532–540. https://doi.org/10.1080/21670811.2019.1591927

Vesselkov, A., Finley, B., & Vankka, J. (2020). Russian trolls speaking Russian: Regional Twitter operations and MH17. *12th ACM Conference on Web Science*, 86–95. https://doi.org/10.1145/3394231.3397898

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. *Companion Proceedings of The 2019 World Wide Web Conference*, 218–226. https://doi.org/10.1145/3308560.3316495

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Who let the trolls out? Towards Understanding State-Sponsored Trolls. *Proceedings of the 10th ACM Conference on Web Science*, 353–362. https://doi.org/10.1145/3292522.3326016