vera.ai: VERification Assisted by Artificial Intelligence

D4.3 - Disinformation Impact and Platform Algorithm Assessments

Project Title	vera.ai
Contract No.	101070093
Instrument	HORIZON-RIA
Thematic Priority	CL4-2021-HUMAN-01-27
Start of Project	15 September 2022
Duration	36 months

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.
The content of this document is © of the author(s) and respective referenced sources. For further information, visit veraai.eu.

Deliverable title	Disinformation Impact and Platform Algorithm Assessments
Deliverable number	D4.3
Deliverable version	V0.1
Previous version(s)	-
Contractual Date of delivery	14.09.2025
Actual Date of delivery	13.09.2025
Nature of deliverable	Report
Dissemination level	Public
Partner Responsible	UvA
Author(s)	Richard Rogers, Kamila Koronska, Varvara Boboc (UvA), Ivan Srba (KInIT), Fabio Gigleitto, Giada Marino, Anwesha Chakraborty (UNIURB), Maria Giovanna Sessa, Raquel Miguel Serrano (EUDL)
Reviewer(s)	Inès Gentil, Francesco Poldi (EUDL), Milica Gerhardt (IDMT)
EC Project Officer	Peter Friess

Abstract	This deliverable advances the objective of enabling the discovery, tracking, and impact measurement of disinformation narratives and campaigns across platforms, modalities, and languages by applying integrated AI and network science methods. Building on the methodological foundations of D4.2, it provides case studies demonstrating the efficiency of these approaches. The report begins with a meta-analysis of disinformation impact indicators and vera.ai tools, highlighting strengths in detecting virality and coordinated inauthentic behaviour

	while identifying gaps in qualitative assessment. It then analyses algorithmic amplification, with TikTok as the platform in focus, exploring systemic risks, democratic implications, and regulatory challenges under the Digital Services Act. Finally, through case studies on TikTok and Meta, the deliverable underscores best practices and recommendations aimed at fostering proactive, evidence-based disinformation monitoring for fact-checkers, researchers, and policymakers.
Keywords	Disinformation assessment, Algorithmic amplification, VLOPs, Coordinated inauthentic behaviour, DSA

Copyright

© Copyright 2025 vera.ai Consortium

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the vera.ai Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

Revision History

Version	Date	Modified by	Comments
V0.1	05/06/2025	Varvara Boboc, Richard Rogers (UvA)	Draft ToC
V0.2	19/07/2025	Varvara Boboc (UvA), Francesco Poldi (EUDL)	Added alignment report 1
V0.3	31/07/2025	Varvara Boboc (UvA), Ines Gentil, Maria Giovanna Sessa, Raquel Miguel Serrano (EUDL)	Added alignment report 2
V0.4	29-31/07/ 2025	Varvara Boboc (UvA), Ivan Srba (KInIT) (as Mosnar et. al 2025 institution representative)	Added TikTok case studies
V0.5	01/08/2025	Giada Marino, Anwesha Chakraborty (UNIURB)	Feedback and comments
V0.6	01-04/08/ 2025	Fabio Giglietto, Giada Marino (UNIURB)	Added Pope Francis, Gambling, Putin Fan Groups and Coordinated Pro-Bolsonaro Facebook Accounts case studies
V0.7	18/08/2025	Kamila Koronska (UvA)	Added Coordinated Networks Spread Political Content on Polish TikTok case study & Feedback and comments
V0.8	18-21/08/ 2025	Varvara Boboc, Richard Rogers (UvA)	Addressed comments, edited overall deliverable
V0.9	10/09/2025	Richard Rogers (UvA), Olga Papadopoulou (CERTH)	Final feedback and comments
V1.0	13/09/2025	Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Deliverable sent to EC

Glossary

Abbreviation	Meaning
DoA	Description of Action
DSA	Digital Services Act
WP	Work Package
CIB	Coordinated Inauthentic Behaviour
CSB	Coordinated Sharing Behaviour
CSDS	Coordinated Sharing Detection Service
FYP/FYF	For You Page / For You Feed
MCL	Meta Content Library
TTPs	tactics, techniques, and procedures
VLOP	Very Large Online Platform

Table of Contents

Index of Tables

Index of Figures

Executive Summary

This deliverable builds on the objective of **enabling the discovery, tracking, and impact measurement of disinformation narratives and campaigns across social platforms, modalities, and languages through integrated AI and network science methods** (SO3). While the previous deliverable, D4.2, presents the methodologies that build the foundation of this objective, the current deliverable provides several categories of case studies that showcase the effectiveness of the methodologies.

Thus, D4.3 commences with a **meta-analysis of disinformation impact indicators** and tools to provide an overview of disinformation measurement assessment and its pitfalls. Within this landscape, vera.ai tools already excel at quantifying virality and detecting non-organic amplification such as Coordinated Inauthentic Behaviour; extending them to qualitative indicators is the next frontier. Overall, progress is real but fragmented, calling for harmonised terminology, richer alert mechanisms, and systematic evaluation of responses.

Afterwards, an outline of **algorithmic amplification** is presented, with a focus on **TikTok** as the core platform of interest. In the wake of TikTok's designation as a Very Large Online Platform (VLOP) under the Digital Services Act (DSA) and its impact on democratic processes such as the Romanian elections, the analysis provides a breakdown of platform risks, research gaps, and policy implications. By defining algorithmic amplification, tackling it as a systemic risk, and exploring the "For You" Page (FYP) and platform design as a risk vector, the report highlights the needs of greater transparency, broader researcher access to data, and a regulatory turn toward amplification as a structural concern under the DSA.

To complete the comprehensive overviews and to exemplify the effectiveness of a series of methodologies presented in D4.2, the rest of the deliverable is dedicated to two categories of **case studies**: (1) detailed examinations of **TikTok's algorithmic amplification** and (2) **investigations of inauthentic coordination**, mainly targeting the Meta platform.

The deliverable concludes with a series of general remarks and specific insights targeted at producing a proactive approach to disinformation monitoring and analysis dedicated to fact-checkers, researchers and policymakers.

1 Introduction

This deliverable deepens the project's SO3, "enabling the discovery, tracking, and impact measurement of disinformation narratives and campaigns across social platforms, modalities, and languages through integrated AI and network-science methods." Whereas D4.2 set out those methods, D4.3 demonstrates their practical value through two meta-analyses and a suite of targeted case studies.

The document opens with a meta-analysis of disinformation impact indicators and the vera.ai toolset and proposed methodologies. Assessing the effective – or potential – impact of mis- and disinformation remains complex. Existing analytical frameworks all try to capture both reach and harm, yet label that dimension variously as "effect," "degree," or "effectiveness," which hampers comparability. They also diverge on which indicators to privilege: sheer online reach, observable offline behaviours, or subtler shifts in attitudes and opinions. Quantitative metrics dominate because they are easier to standardise, while qualitative and psychological aspects are acknowledged but underdeveloped, typically imported piecemeal from the social sciences. Methodological rigour likewise varies: some models are cautiously empirical, others more speculative. Only two out of six frameworks tackled in the meta-analysis embed warning systems that flag high-risk situations. Furthermore, almost no framework evaluates how countermeasures alter impact, leaving the attacker's perspective and defender responses largely uncharted. Within this landscape, vera.ai tools already excel at quantifying virality and detecting non-organic amplification such as Coordinated Inauthentic Behaviour (CIB). Additionally, mixed methods protocols for assessing the impact of disinformation developed by vera.ai partners – the Post-truth Space Mapping Technique Protocol and the CIB Typology Development Technique – enable a more systematic investigation of disinformation dynamics. Overall, current methodologies show progress but remain fragmented. Greater alignment on terminology, stronger integration of qualitative and psychological factors, improved alert mechanisms, and an emphasis on countermeasures are all necessary to advance impact assessment in disinformation and influence operations.

The second meta-analysis investigates how algorithmic amplification on TikTok's "For You[1]" Page is defined and understood in current research, despite the lack of access to proprietary platform data. A review of six recent academic studies suggests that amplification is driven mainly by implicit signals like watch time, leading to rapid personalisation that steers users into narrow content loops. This raises concerns about reduced content diversity, limited agency, and systemic risks linked to the amplification of harmful clickbait and divisive content, including mis- and disinformation. An additional risk is the exploitation of algorithmic amplification mechanisms by malign actors for purposes of Foreign Information Manipulation and Interference (FIMI), electoral disruption, or broader societal destabilisation. TikTok's DSA risk assessment reports offer only limited insights into algorithmic risks but remain content-focused and provide little insight into how amplification operates or is mitigated. The report therefore urges greater transparency, broader researcher access to data, and a shift in regulatory focus toward addressing amplification as a structural concern in compliance with the DSA.

[1] https://support.tiktok.com/en/getting-started/for-you

A series of case studies elaborates on the overarching topics of the two meta-analyses, providing a multifaceted view on the strengths and limitations of current disinformation detection methods and coordinated inauthentic behaviour assessment.

1.1 Deliverable Structure

To translate these insights into practice, the remainder of D4.3 is devoted to two clusters of case studies: (1) three detailed examinations of TikTok algorithmic amplification phenomena and (2) four investigations of inauthentic coordination, mainly targeting the Meta platform. Each case operationalises the methodologies introduced in D4.2, surfacing research gaps and concrete policy implications. The deliverable concludes with best practices and recommendations designed to foster a proactive stance among fact-checkers, researchers, and policymakers, moving the field from reactive debunking toward anticipatory, evidence-informed governance of disinformation risks.

1.2 Key Performance Indicators (KPIs)

The project's co-creation activities significantly exceeded the established impact targets. For **co-creation workshops**, the target was set at *more than three*. This was surpassed with a total of **eight workshops**, including *three Winter Schools, three Summer Schools, the Bellingcat Hackathon, and a Verification Sprint before the EU Parliamentary Election*.

In terms of the number **of co-creation workshop participants**, the target was set at *more than 100*. The achieved result exceeded expectations, with an **approximate total of 1,400 participants**. Each Winter School attracted over *300 participants*, while each Summer School hosted over *150 participants*. In addition, the *Bellingcat Hackathon brought together 20 participants*, and the *Verification Sprint involved 18 participants*.

Finally, regarding the **number of stakeholder groups involved in the workshops**, the target was to engage *more than four groups*. This was successfully expanded to a total of **12 stakeholder groups**, covering a wide range of expertise and perspectives. The groups included *Bellingcat, AI Forensics, Adapt Institute, Demagog, Lakmusz, Delfi, Post-X Society, SCPS (Sustainable Cooperation for Peace & Security), Alliance for Securing Democracy (German Marshall Fund), ISD (Institute for Strategic Dialogue), NRC Handelsblad, and Agence France-Presse (AFP)*.

Taken together, these results demonstrate strong engagement, diversity, and scalability across workshops.

In revising the KPI on *agreement between disinformation impact assessments and expert ratings* related to the advancement of **multidimensional methodologies for assessing the impact of disinformation** – originally operationalised as a Pearson correlation > 0.6, it became evident that the challenge lies less in statistical concordance than in the absence of consensus on what constitutes "impact" in the first place. As argued throughout the deliverable, disinformation impact assessment is a multidimensional problem, complicated by ongoing debates around authenticity measures within coordinated inauthentic behaviour (CIB) frameworks.

To address this, our approach moved beyond a narrow correlation metric and instead emphasised the co-development of methodologies with domain experts. During the 2024 Digital Methods Initiative Summer School in Amsterdam[2], a **dedicated expert group tested and demonstrated the effectiveness** of characterising **coordination** through engagement patterns, specifically using degree distribution analysis. This collaborative effort produced a validated typology of coordinated networks (i.e., media, influence operations, grassroots, advertising, and exploited public groups) each with distinct modes of amplification and reach. By reframing "degree of coordination" as an **impact measure**, we **aligned the KPI with both expert practice and real-world observability, thus ensuring it remains both robust and meaningful**. Importantly, this methodology has already been applied to networks flagged by Vera AI Alerts, showing clear potential for practical deployment and external validation. The advancement of this approach, alongside complementary innovations such as the **Post-truth Space Mapping Protocol** and **the CIB Typology Development Technique**, strengthens the multidimensional framework introduced in the deliverable, while its scholarly underpinning is documented in forthcoming publications: "Coordinated Inauthentic Behaviour on Facebook? A Typology of Manufactured Attention," Platforms & Society, SAGE." (Rogers & Righetti, 2025) and "Post-Truth Spaces: Studying Authenticity and Influence on the Internet" (Rogers & Koronska, 2025).

[2] https://www.digitalmethods.net/Dmi/SummerSchool2024

2 A Meta-analysis of Disinformation Impact Indicators and vera.ai Tools: Achievements, Gaps, and Opportunities

2.1 The Challenges of Measuring Impact: Navigating "the Illusion of Metrics" and Inconsistent Definitions

Assessing the effective or potential impact of misinformation and disinformation remains one of the most significant and widely recognised challenges in the field. Technology has fuelled what influence operations expert Thomas Rid describes as the "seductive illusion of metrics" in his book *Active Measures* (2020). In an interview with Kate Starbird, he adds that "measuring the actual impact of trolling and online influence campaigns is probably impossible (…) but the difficulty of measuring impact doesn't mean that there isn't meaningful impact."

This raises fundamental questions: How should we define impact, and how can we assess whether disinformation or influence campaigns had an impact or pose a threat of potential impact? Should this impact or risk be measured by the online reach of specific content, real-world offline actions, or shifts in public opinion, attitudes, and behaviours?

One of the biggest hurdles in researching the effective or potential impact of influence campaigns and disinformation lies in the very definition of the concept itself, which can be ambiguous and lacks consensus due to various methodologies and frameworks proposing different interpretations.

On one hand, it may refer to the actual effectiveness of an information operation in driving actions or attitude changes, or influencing the opinions or behaviours of those exposed to it. This approach prioritises offline consequences, evaluating how online disinformation translates into real-world harm. This includes analysing calls to action or instances where false narratives trigger behavioural changes, including harmful conduct such as individuals drinking bleach[3] to cure COVID-19 or the Pizzagate conspiracy[4] leading to the harassment of restaurant employees. This research approach often requires proving a direct cause-and-effect relationship between disinformation and changes in belief or behaviour, which is normally challenging, as well as a direct interaction with human subjects to assess those changes.

On the other hand, impact is often used more loosely to refer to the potential influence of an operation, typically inferred from digital traces such as engagement metrics or reach. In this sense, some researchers define the risk of impact in terms of outreach and virality, emphasising the number of people exposed to a particular content. Others focus on engagement, measuring how actively people interact with disinformation. From this perspective, the greater the circulation of a false narrative across platforms, communities, and languages, the higher the threat of impact, as it can, for instance, fuel hateful and discriminatory narratives against specific groups. This research approach is based on observable elements within information environments.

[3] https://www.forbes.com/sites/nicholasreimann/2020/08/24/some-americans-are-tragically-still-drinking-bleach-as-a-coronavirus-cure/

[4] https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

In general terms, the analysed frameworks do not explicitly distinguish between effective and potential impact, except the Breakout Scale, which acknowledges that metrics on social media platforms are just an approximation of impact, and the Impact-Risk Index, which, as its name indicates, studies the potential influence of a piece of disinformation.

Most existing approaches frame impact negatively, focusing on the harm caused by disinformation campaigns. However, impact can also be understood positively through the effectiveness of countermeasures deployed to combat the harm, outreach, and overall effects of disinformation. In this sense, measuring impact helps determine not only the potential risks posed by a hoax or influence campaign but also the success of interventions in mitigating its reach and harm.

Thus, an effective impact assessment framework should capture both:

1. The adverse effects of a disinformation campaign.
2. The positive effects of countermeasures, reflected in a measurable decline in disinformation impact indicators.

This study contributes to the evolving understanding of the impact of disinformation by mapping current approaches and identifying key indicators used for assessment. In this regard, the vera.ai project has developed tools that can significantly support the evaluation of these indicators. By linking these tools directly to relevant metrics, we aim to maximise their utility while also identifying existing gaps and limitations. This approach not only provides a roadmap for future research but also highlights emerging challenges and opportunities for researchers and developers seeking to enhance impact measurement methodologies.

2.2 A Compilation of Existing Frameworks: Selection Criteria and Focus

2.2.1 Selection Criteria

With these objectives in mind, we begin by compiling an inevitably non-exhaustive list of key frameworks used to study disinformation and influence operations. We then critically examine how these frameworks define and, in some cases, measure the effective or potential impact of disinformation, as well as the indicators they employ for assessment. This analysis enables us to conduct a comparative evaluation, identifying the similarities, differences, and gaps between existing methodologies.

Finally, we explore how the tools developed within the vera.ai project can support the measurement of these impact indicators. By linking these tools to relevant metrics, we aim to maximise their utility while also identifying existing gaps and limitations, providing a roadmap for the future of impact measurement.

When compiling the frameworks for analysis, we focus on comprehensive study methodologies that examine the entire lifecycle of hoaxes, disinformation campaigns, and influence operations – and thus consider impact as one aspect within a broader analysis – as well as impact-specific methods that take a more targeted and in-depth approach.

2.2.2 Comprehensive Frameworks to Study the Disinformation Lifecycle

The following frameworks focus on the entirety of the disinformation cycle:

1. ABCDE Framework (Pamment , 2020)[5]
2. Disarm Red Framework / Disarm Blue Framework[6]
3. CIB Detection Tree[7]

While we concentrate on the entirety of the disinformation cycle, we also would like to acknowledge other frameworks that address specific aspects of disinformation but not the whole process, such as "Directing responses against illicit influence operations[8]" (D-RAIL), "Attributing information and influence operations" (Pamment & Smith, 2022) or "DETERRENCE: Proposing a more strategic approach to countering hybrid threats" (Vytautas, 2020).

2.2.3 Frameworks Specifically Focused on the Impact of Disinformation

These include:

1. Breakout Scale[9]
2. Impact-Risk Index[10]
3. Response-Impact Framework[11]

In addition to these disinformation-focused frameworks, social science methodologies are used to study behaviour and behaviour change, such as the COM-B Model[12] for behaviour change (identifying Capability, Opportunity, and Motivation as the key drivers of Behaviour) (West & Michie, 2020) or the use of control groups or A/B testing[13], a method for comparing two versions of a single variable to evaluate which is more effective. Since these methodologies are not explicitly designed for analysing information dynamics, they cannot be subjected to the same critical comparative analysis as the others.

[6] https://disarmframework.herokuapp.com
[7] https://www.veraai.eu/posts/report-visual-assessment-of-cib-in-disinfo-campaigns
[8] https://www.disinfo.eu/wp-content/uploads/2024/08/20240818_D-RAIL.pdf
[9] https://carnegieendowment.org/research/2020/07/the-eus-role-in-fighting-disinformation-taking-back-the-initiative?lang=en
[10] https://www.disinfo.eu/wp-content/uploads/2022/09/20220610_IndexImpactAssessment_Final-1.pdf
[11] https://www.disinfo.eu/publications/beyond-disinformation-countermeasures-building-a-response-impact-framework/
[12] https://thedecisionlab.com/reference-guide/organizational-behavior/the-com-b-model-for-behavior-change
[13] https://hbr.org/2017/06/a-refresher-on-ab-testing

2.3 A meta-analysis of Selected Frameworks: Definitions, Indicators, Speculation and Alerts

2.3.1 Terminology and Definitions

First, we examined the frameworks in terms of their coverage and approach to measuring the impact of disinformation.

- Does the framework cover the impact of disinformation?
- If affirmative, in which sense: outreach, harm, effect of countermeasures?

Table 1 presents the terminology of "impact" in the different frameworks.

Table 1 Terminology of "impact" in the different frameworks

Framework	Covers The Topic	Impact As Outreach	Impact As Harm	Impact As Countermeasures' Effect
ABCDE Framework	X	X	X	
Disarm Red/Blue Framework	X	X	X	
CIB Detection Tree	X	X	X	
Breakout Scale	X	X	X	
Impact-Risk Index	X	X	X	
Response-Impact Framework	X			X

1. ABCDE Framework (Pamment, 2020)

This framework, developed in the context of the EU's intention to tackle disinformation and foreign influence operations, breaks down an influence or disinformation campaign into five key elements: Actor, Behaviour, Content, Degree, and Effect. Therefore, impact is considered in two distinct ways.

- Impact as outreach. In this context, potential impact refers to the extent to which the content spreads, captured under "**Degree**" (D). It is measured using indicators such as cross-platform distribution, media coverage, and amplification in multiple languages.
- Impact as harm. Here, effective impact is understood as the harm inflicted, identified through its "**Effect**" (E). It is measured by its influence on polarisation, the discrediting of institutions, risks to public health and safety, threats to freedom of thought and expression, and security risks at personal, organisational, and national levels.

2. Disarm Red Framework / Disarm Blue Framework

The Red Framework, developed by the DISARM Foundation, focuses on the Tactics, Techniques, and Procedures (TTPs) employed by disinformation actors across all campaign phases: Plan, Prepare, Execute, and Assess. In this approach, impact is viewed from the perspective of the disinformer, referring to the campaign's **effectiveness**, both in terms of outreach and its ability to influence attitudes and behaviour.

- Impact as outreach. In terms of outreach, the relevant TTPs are "T0133: Measure Effectiveness" and "T0134: Measure Effectiveness Indicators (or KPIs)". This tactic emphasises the importance of establishing key performance indicators in advance, allowing the campaign's success to be evaluated afterwards based on metrics such as message reach and social media engagement.
- Impact as harm. Regarding harm, campaign effectiveness includes monitoring and evaluating changes in audience knowledge, attitudes, or behaviour.

Table 2 TTPs included in the Disarm Red Framework related to impact measures

TTPs	Focus	Description
T0134 Measure effectiveness indicators (KPI)		
T0134.001	Message reach	Monitor and evaluate message reach in misinformation incidents.
T0134.002	Social media engagement	Monitor and evaluate social media engagement in misinformation incidents.
T0133 Measure effectiveness		
T0133.001	Behaviour changes	Monitor and evaluate behaviour changes from misinformation incidents.
T0133.002	Content	Measure current system state with respect to the effectiveness of campaign content.
T0133.003	Awareness	Measure the current system state with respect to the effectiveness of influencing awareness.
T0133.004	Knowledge	Measure the current system state with respect to the effectiveness of influencing knowledge.
T0133.005	Action/attitude	Measure current system state with respect to the effectiveness of influencing action/attitude.

Table 2 illustrates the TTPs included in the Disarm Red Framework related to impact measures. The Blue Framework outlines the responses given by the counter-disinformation community – i.e., actions designed to counter the attacker's strategies directly. Nonetheless, it does not elaborate on counter-strategies or

explain how to assess their impact. That said, the framework is expected to be revised, which will likely bring greater clarity on these aspects.

3. CIB Detection Tree

The CIB Detection Tree, developed by EU DisinfoLab in 2021 and updated in 2024 as part of the vera.ai project, outlines four analytical branches for detecting Coordinated Inauthentic Behaviour (CIB) within the context of influence or disinformation campaigns. Notably, one of these branches is entirely dedicated to "**Impact Assessment**", which is primarily understood as outreach in a broad sense, encompassing not only outreach but also engagement and interaction metrics. Moreover, it also considers the potential impact on the targets and the polarising effects of the content.

4. Breakout Scale

The Breakout Scale, developed by Ben Nimmo, is one of the earliest frameworks specifically focused on impact and was designed to evaluate influence operations. Its purpose is to enable operational researchers to assess and compare the likely impact, both in terms of outreach and harm, of different influence operations in real time, using measurable and replicable evidence as the basis for evaluation.

While acknowledging that "it is not practically possible to measure sentiment change, and thus to arrive at a direct assessment of impact", it states that "if an influence operation is to change the beliefs or behaviour of a community, it has to be able to land its content in front of that community first, and the way a message passes from one community to another can be tracked and measured" (Nimmo, 2020).

- In this regard, potential impact is intended as outreach "on social media, in the mainstream media, and real life". While acknowledging that metrics on social media platforms are an approximation of impact, this framework enhances the outreach-based approach by incorporating insight and analysis into how an influence operation expands beyond its initial target community, offering a broader understanding of its spread and potential reach.
- However, effective impact is intended as harm, especially when an influence operation reaches the highest category of the scale, "category six", i.e., when "it triggers a policy response or some other form of concrete action, or if it includes a call for violence".

5. Impact-Risk Index

While the Breakout Scale is designed to be applied to influence operations, the EU DisinfoLab Impact-Risk Index offers an approach to assess the potential impact of a single hoax. The method involves a simple list of eight indicators, whose scores are translated into a final scale measuring the low, medium, high, or alarming impact risk. The framework understands the potential impact in terms of both outreach and harm.

- Impact is intended as the impact on diffusion, related to the virality, but also with a special focus on engagement, of a single disinformative content.
- Impact is intended as the impact on harm when assessing if a hoax contains a "call to action", which indicates a possible offline effect of the disinformative piece of content.

6. Response-Impact Framework

The EU DisinfoLab's Response-Impact Framework offers a fundamentally different perspective, designed not to measure the impact of disinformation itself, but rather the effectiveness of the countermeasures implemented to combat it.

We include this framework here because it addresses a dimension often overlooked in other methodologies – i.e., the impact of responses. In this model, the greater the effectiveness of countermeasures, the lower the impact of the disinformation campaign will be. These countermeasures aim to both limit outreach and amplification (through distribution-related interventions) and mitigate harm (via infrastructure-targeted actions, sanctions, and legal responses).

2.3.2 Qualitative v. Quantitative Techniques and Indicators

Having explored various approaches to defining impact, we now turn to how these methodologies actually measure it – that is, the indicators they use. To facilitate analysis, we have divided the suggested techniques into two broad categories: quantitative and qualitative.

- Quantitative techniques: These refer to methods that produce measurable, numerical data – for example, the number of media outlets publishing a hoax, the number of platforms on which it appears, or the number of likes and shares a post receives. Most of these indicators align with impact as outreach or engagement in its most basic form, but can also take the form of A/B testing methods.
- Qualitative techniques: These involve methods that do not easily translate into numerical values, making them more difficult to standardise across studies, such as surveys, focus groups, and expert interviews. These approaches often aim to capture more nuanced, context-driven insights into the effects of disinformation.

However, the line between quantitative and qualitative methods is often not clear-cut. For instance, the Breakout Scale considers amplification by a public figure as a key factor – something that could be interpreted both qualitatively (through sentiment analysis) and quantitatively (in terms of views to a particular post). Similarly, the Impact-Risk Index attempts to quantify elements such as calls to action by assigning them numerical scores, thereby increasing or decreasing the perceived level of risk.

To better understand the practical application of each framework, we examine the techniques they propose or endorse for measuring effective or potential impact – and, crucially, if they provide clear guidelines or methodologies for applying these techniques in practice (see Table 3). We also look at how these frameworks incorporate social and psychological dimensions and include a final note highlighting additional methods that do, even if they are not part of the core list analysed. Understanding the types of indicators used (and how clearly, they are defined) helps assess the operational readiness and practical utility of each framework, especially for researchers, policymakers or tool developers seeking to evaluate disinformation impact.

- Does the framework require the use of quantitative techniques to measure impact?
- Does the framework require the use of qualitative techniques to measure impact?
- Does the framework provide clear guidelines on how to carry out such measurements?

Table 3 Quantitative and qualitative techniques in the measurement of impact

Framework	Quantitative techniques	Qualitative techniques	Guidelines
ABCDE Framework	X	X	
Disarm Red/Blue Framework	X	X	
CIB Detection Tree	X	X	
Breakout Scale	X	X	X
Impact-Risk Index	X	X	X
Response-Impact Framework	X	X	

Table 4 Detailed indicators of impact in the different frameworks

FRAMEWORK	PLATFORM-RELATED INDICATORS			CONTENT-RELATED INDICATORS		MEDIA	NUMBER OF COMMUNITIES	TYPES OF HUMAN AMPLIFIERS		BELIEF CHANGES INDICATORS		BEHAVIOUR CHANGES-OFFLINE ACTIONS (policy responses/calls to violence)	
	# of platforms	Reach (views)	Engagement (shares, reactions)	Language	Format	Fringe or mainstream media		Public figures	Recurrent disinformers	Sentiment analysis indicating polarisation, institutional discredit	Limitation/ suppression of freedom of thought and speech	Public health or safety risk	Limitation/ suppression of freedom of thought and speech
ABCDE Framework	X			X		X				X	X	X	X
Disarm Red/Blue Framework	X	X	X							X		X	
CIB Detection Tree	X	X	X	X		X		X		X	X		X
Breakout Scale	X					X	X	X				X	
Impact-Risk Index	X	X	X	X	X	X		X	X			X	
Response-impact framework	X					X				*positive outcome		*positive outcome	

Table 4 summarises the indicators of impact in the difference frameworks.

1. ABCDE Framework (Pamment, 2020)

The ABCDE Framework recommends the use of both quantitative and qualitative techniques, yet it remains somewhat descriptive. Overall, it does not provide clear guidance, best practices, or a defined methodology for putting these measurements into practice.

- Key elements for measuring "Degree" – such as cross-platform distribution, virality metrics, media coverage, and multilingual amplification – primarily rely on quantitative techniques, involving measurable and comparable data.
- In contrast, key elements for measuring "Effect" – including polarisation, the discrediting of institutions, risks to public health or safety, threats to freedom of thought and expression, and personal, organisational, or national security risks – require a more qualitative approach, as they involve complex, context-dependent assessments that are less easily quantified.

- Indicators of Degree: platforms, media, languages.
- Indicators of Effect: polarisation; discredited institutions; public health or safety risk; freedom of thought and expression; risks to personal, organisational, and national security.

2. Disarm Red Framework / Disarm Blue Framework

The DISARM framework includes specific TTPs for measuring impact in the campaign's "Assess" phase.

- The TTPs related to "T0134: Measure Effectiveness Indicators (or KPIs)" – based on media reach and engagement – suggest that quantitative techniques are needed.
- The ones related to measuring the campaign's "effectiveness" – e.g., evaluating changes in audience knowledge, attitudes, or behaviour – point to qualitative methods.

However, the framework does not provide specific guidance on how to conduct these measurements.

- Indicators of reach: message reach; social media engagement.
- Indicators of "effectiveness"/harm: changes in behaviour, knowledge, awareness, and attitude.

3. CIB Detection Tree

The application of the CIB Detection Tree involves the use of both quantitative and qualitative techniques.

- To assess virality, the framework relies on quantitative indicators, including the volume of likes and reshares, media amplification, and the presence of content across multiple social media platforms. It emphasises the number of users reached over time, even considering it more significant than interaction metrics like likes or shares.

- At the same time, the framework incorporates qualitative indicators that require the use of qualitative research techniques. These include the presence of polarising content, the targeting of specific groups, amplification by public figures or influencers, sentiment analysis, and the suppression of dissenting opinions in favour of dominant narratives.

- Indicators of reach: multiple social media platforms, the number of users reached over time, and interaction metrics such as likes or shares.
- Indicators of harm: polarising content, specific targets, amplification by public figures or influencers, sentiment analysis, and suppression of dissenting opinions that favour dominant narratives.

4. Breakout Scale

The application of the Breakout Scale involves both quantitative and qualitative methods. While it incorporates some numerical data, the approach places greater emphasis on qualitative analysis. According to the author, platform metrics alone are insufficient and potentially misleading indicators of the actual impact of influence operations, especially when they spread across platforms. In addition, metrics fail to account for any artificial amplification, the size of the initially targeted audience segments, or the actual exposure time each user receives.

- The Breakout Scale relies on quantitative measures such as the number of platforms, media outlets, and communities targeted to assess outreach.
- However, the focus is primarily on understanding how the breakout spreads across different communities. In addition, evaluating harm requires qualitative techniques, focusing on key indicators such as offline actions and real-world consequences.

- Quantitative indicators: number of platforms, media outlets, and targeted communities.
- Qualitative indicators: relevant human amplification; offline actions such as policy responses and calls for violence.

7. Impact-Risk Index

The Impact-Risk Index applies both quantitative and qualitative methods, but prioritises the former. While the Breakout Scale is intended for broader and more complex influence operations, the Impact-Risk Index is tailored to assess individual hoaxes. This narrower focus on specific and concrete units enables a higher degree of quantification.

- Key elements to be measured quantitatively include exposure and engagement indicators, as well as the number of platforms, languages, and media outlets involved.
- At the same time, the index also incorporates qualitative aspects, such as the format of the hoax, the prominence of individuals sharing the content, the identification of persistent disinformation transmitters, and any associated calls to action.

- Indicators of reach: Indicators used to assess diffusion include both engagement metrics (such as shares and reactions) and exposure metrics, including the number of views. Additional quantitative indicators are the number of platforms and languages involved, as well as the extent of media outreach.
- On the qualitative side, the analysis considers the type of actors disseminating the content, such as public figures or recurrent disinformation purveyors, the format of the content, and the presence of calls to action.

8. Response-Impact Framework

- In this context, impact refers to the effectiveness of the countermeasures in curbing the disinformation campaign. Certain elements require quantitative measurement methods, particularly those assessing the amplification and distribution of threat actors. This is evaluated by observing a demonstrable decrease in the campaign's reach, using the same indicators – such as cross-platform amplification and engagement metrics – but through a comparative analysis with prior data.
- Impact is also understood as promoting positive outcomes within the defenders' community, including increased awareness of the threat, successful attribution of the campaign, initiating further actions, and overall improvements in societal resilience. These aspects would require qualitative methods to assess effectively.

- Indicators of reach: cross-platform amplification and exposure/engagement metrics. (in the sense of a decreasing impact).
- Qualitative indicators reveal positive elements, including increased awareness, improved attribution accuracy, increased action, and enhanced societal resilience.

2.3.3 Speculation Level and Alert Systems

After a detailed analysis of the frameworks and their approaches to measuring impact, it is necessary to include a disclaimer regarding the accuracy of impact assessment and the degree of speculation involved in drawing conclusions. A key question is: to what extent is the evidence within each framework robust enough to support its conclusions? To address this, we assess the level of speculation, classified as low, medium, or high, based on the presence or absence of clearly defined indicators and the strength of the evidence required to support findings.

- Low: The framework indicates specific elements to measure impact and provides concrete indicators to draw conclusions.
- Medium: The framework indicates specific elements to measure, but relies primarily on subjective assessment to draw conclusions.
- High: The framework does not specify key elements to measure and lacks evidence to assess its impact.

At the same time, we also consider whether the frameworks incorporate any form of alert or early warning system that could help investigators detect and anticipate the potential threats posed by specific campaigns, incidents, or hoaxes.

- What is the level of speculation about the impact (or is the assessment linked to concrete measures)?
- Does the framework include an alert system?

Table 5 presented the speculation level of each framework and labels the frameworks including an alert system.

Table 5 Frameworks' speculation level

Framework	Speculation Level			Alert System
	Low	Medium	High	
ABCDE Framework		X		
Disarm Red/Blue Framework		X		
CIB Detection Tree		X		
Breakout Scale	X			X
Impact-Risk Index	X			X
Response-Impact Framework			X	

1. ABCDE Framework (Pamment, 2020)

The ABCDE Framework identifies elements that define the effective or potential impact of a campaign. Still, it does not provide guidance on how to carry out measurements or draw conclusions from an impact assessment. For this reason, we assess a medium level of speculation.

2. Disarm Red Framework / Disarm Blue Framework

The DISARM framework also highlights elements that define a campaign's effectiveness, and stresses the importance of setting KPIs. However, it does not establish specific models or methodologies and leaves out a responsibility for drawing conclusions about the campaign initiators. As a result, we consider it to have a medium level of speculation.

3. CIB Detection Tree

The CIB Detection Tree offers concrete instructions for measuring effective or potential impact and provides supporting tools. Nevertheless, it lacks a clear methodology for interpreting results or drawing conclusions, which is why we classify it as having a medium level of speculation.

4. Breakout Scale

The Breakout Scale presents a specific methodology for measuring effective or potential impact by listing relevant indicators while carefully interpreting metrics. Due to its structured yet conservative approach, we assess a low level of speculation.

9. Impact-Risk Index

The Impact-Risk Index is highly detailed in listing measurable indicators, even including thresholds, scoring systems, and clear guidance for assessing impact. For this reason, it is considered to have a low level of speculation.

10. Response-Impact Framework

Finally, the Response-Impact Framework takes a different approach and, as such, offers a limited perspective on the impact of campaign distribution. It does not provide specific guidance on how to carry out measurements or draw conclusions. For this reason, its level of speculation can be considered high.

A warning system is essential for determining when heightened monitoring and precautionary measures are needed in response to a campaign or hoax with a potentially greater impact. It also helps to prioritise and focus on incidents that present a higher level of risk.

Among the frameworks reviewed, only EU DisinfoLab's Impact-Risk Index and Ben Nimmo's Breakout Scale include systems that define thresholds or indicators beyond which the threat level significantly increases.

2.4 Learning outcomes: Similarities, Differences, and Drawbacks

2.4.1 Commonalities and Divergences

All the frameworks address the concept of "impact" in some form, usually referring to both reach and harm. However, they differ in terminology, emphasis, and focus.

- Terminologically, various concepts are employed, including "degree", "effect", "effectiveness", "changes", and "impact".
- Regarding emphasis, the ABCDE and DISARM frameworks adopt a relatively balanced approach, considering both dimensions. In contrast, the CIB Detection Tree and the Breakout Scale place greater weight on reach, though the latter does so with a notably sceptical stance toward metrics. The Impact-Risk Index is heavily focused on reach and engagement indicators, whereas the Response-Impact Framework offers a distinct perspective by looking at the effectiveness of countermeasures.

Many elements are shared across the frameworks, particularly regarding platform amplification, yet the approach and level of detail vary significantly. For instance, while the ABCDE and Response-Impact Frameworks briefly mention cross-platform amplification, others like DISARM, the CIB Detection Tree, and the Impact-Risk Index offer more detailed insights, including specific KPIs related to reach and engagement. Some frameworks focus on only one of these aspects, such as the Breakout Scale, which refers solely to the number of platforms involved.

The content itself is rarely examined in depth. However, a few frameworks, such as ABCDE, highlight the importance of language, while the CIB Detection Tree and Impact-Risk Index go further by identifying content format as an amplification factor, suggesting that a greater variety of formats increases the opportunity for reach and impact.

Media amplification is addressed by nearly all frameworks, typically assessing whether the disinformation remains within fringe outlets or spills over into mainstream media.

Notably, the Breakout Scale introduces a unique indicator: i.e., the number of distinct communities reached by a piece of content. This metric blends quantitative and qualitative analysis, departing from traditional volume-based measurements.

The role of human amplifiers is explicitly recognised in the CIB Detection Tree, Breakout Scale, and Impact-Risk Index. The latter also highlights the presence of recurrent disinformers as a key amplification factor.

Regarding psychological and social indicators, most frameworks acknowledge the danger of changing beliefs and attitudes, institutional distrust, polarisation, and threats to freedom of thought and expression. However, some frameworks, like ABCDE and the CIB Detection Tree, provide more granular analysis of these effects.

Offline actions are referenced across all frameworks, though some differentiate more clearly between specific risks (e.g., threats to public health or safety) and broader societal consequences. Calls to action are considered key indicators in both the Impact-Risk Index and the Breakout Scale.

Finally, it is essential to highlight the Response-Impact Framework's positive approach, which focuses on the beneficial effects of countermeasures. These include mental and behavioural changes such as greater public awareness, increased resilience, and strengthened capacity among the defender community.

2.4.2 Limitations, Gaps, and Challenges

While all frameworks recommend using both quantitative and qualitative techniques for impact assessment, only the Breakout Scale and the Impact-Risk Index provide clear, actionable guidance on how to conduct these measurements. These two offer specific indicators, scoring systems, and thresholds, whereas the others remain more general or conceptual, leaving methodological implementation to the user's discretion.

However, it is crucial to consider the different natures and objectives of the frameworks under review. On one hand, there are comprehensive methodologies that examine the entire lifecycle of hoaxes, disinformation campaigns, and influence operations that do not address impact in great detail. On the other hand, there are specialised frameworks that focus exclusively on measuring impact, providing more targeted and in-depth insights.

A few frameworks directly reference social and psychological factors, such as DISARM, acknowledging psychological dimensions and changes in attitude. However, these aspects are addressed only briefly and conceptually, with no clear methodologies or practical guidelines provided for studying or measuring them effectively. This highlights a broader gap in integrating these considerations across existing frameworks.

Given the inherent difficulty in establishing a direct cause-and-effect relationship between a disinformation campaign and its impact, all the frameworks reviewed involve some speculation. Nonetheless, frameworks like the Breakout Scale and the Impact-Risk Index adopt a more cautious and evidence-based approach, maintaining a relatively low level of speculation. In contrast, others are more general and interpretative, leading to inconsistent reliability when drawing conclusions about impact.

Absence of alert mechanisms to prioritise responses. Most frameworks do not incorporate a dedicated system of alert mechanisms to identify high-risk cases and support the prioritisation of responses. Only two provide warning mechanisms designed to help concentrate efforts with the most significant potential impact.

Most frameworks do not include a positive dimension of impact assessment – namely, evaluating the effectiveness of countermeasures in reducing the reach and harm of disinformation or influence campaigns. This perspective remains largely overlooked in current methodologies.

2.5 vera.ai: How the Project's Tools can Measure Impact

The tools and methodologies developed under the vera.ai project can play a significant role in impact assessment. This section offers insights into how to leverage and integrate them into existing frameworks. This exercise raises awareness of new and expanded use cases while also identifying gaps, challenges, and opportunities for the project's future development.

Tools developed within the vera.ai project:

- **Databases of debunks**. Several tools developed within the vera.ai project – including the **Database of Known Fakes (**DBKF[14]**)** by ONTO – offer valuable support in detecting disinformation, thereby contributing indirectly to assessing impact. For example, if a narrative, image, meme, or video has already been debunked, these tools enable the tracing of the content's amplification loop across other countries, languages, platforms, or information environments.
- Similarly, other disinformation detection tools provide insights on a hoax's previous outreach, such as the reverse image search feature in the Verification Plugin[15]. For instance, by identifying where a manipulated or doctored image has previously appeared, the tool can reveal patterns of cross-platform amplification, helping to trace the spread of the hoax.
- The Vera AI Alerts implements on Meta platforms at a global scale a 9-step workflow designed to track coordinated behaviour on social media (Giglietto et al., 2023). A detailed description of the Vera AI Alerts implementation is available in Section 3.3[16] of deliverable D4.2 Coordinated sharing behaviour detection and disinformation campaign modelling methods.
- **SNA analytics in the** Verification Plugin. The plugin features two data analysis tools: Twitter SNA for X (working with past data up to July 2023) and Facebook and Instagram CSV analysis (utilising CSV files exported from CrowdTangle, which became unavailable since August 2024). The tool enables data collection analysis related to reach and engagement aspects, as well as the detection of actors playing a central role in the distribution scheme, thereby assisting in identifying potential public figures or recurrent disinformation amplifiers who echo the content.
- **XNetwork in the** Verification Plugin. The XNetwork tool enables cross-network queries, helping to identify if the content has been amplified on different platforms. It provides data from X, Facebook, Instagram, YouTube, Reddit, 4chan, LinkedIn, VK, and TikTok, subject to the limitations of the Google Search feature.

[14] https://dbkf.ontotext.com/#!/searchViewResults?orderBy=date&page=1
[15] https://chromewebstore.google.com/detail/fake-news-debunker-by-inv/mhccpoafgdgbhnjfhkcmgknndkeenfhe?pli=1
[16] https://veraai-cms-files.s3.eu-central-1.amazonaws.com/D4_2_V1_0_013beb034d.pdf#page=43.49

- **CooRnet**[17] (Giglietto et al., 2020). This package detects coordinated link-sharing behaviour (CLSB) based on a set of URLs, and maps the network of entities involved. CLSB refers to the coordinated activity of a Facebook network of public pages, groups, and verified profiles that repeatedly share the same news articles within a short time frame. CooRnet helps identify key amplifiers and contributes to assessing cross-platform reach and coordination.
- **CooRTweet R Package**[18] (Righetti & Balluff, 2025) Coordinated Networks Detection on Social Media: Detects a variety of coordinated actions on social media and outputs the network of coordinated users along with related information.
- **CooRTweet interface by Athens Technology Center.**[19] The Coordinated Sharing Detection Service is a tool designed for data journalists, fact-checkers, researchers, and other practitioners to analyse and visualise social media activity through patterns of coordinated sharing. Key features include dynamic network visualisations, exploration of coordinated content, and downloadable outputs for further use. By uploading their data, users can uncover networks of coordinated behaviour, gaining insights for research and monitoring.

For CooRnet, CooRTweet and the CoorTweet interface we have developed an impact assessment methodology based on a coordinated inauthentic behaviour typology developed by the University of Amsterdam and the University of Urbino. We found that the CooRnet and CooRTweet methods are effective in discovering impactful coordination campaigns by adversarial actors but also organic amplification by what one could term authentic actors, like global media groups.

Multidimensional disinformation impact assessment methodologies developed with the vera.ai project:

- **CIB Detection Tree.**[20] This EU DisinfoLab methodology framework was further expanded within the context of the vera.ai project by streamlining its four original assessment branches (Coordination, Source, Impact, Authenticity) into one document, critically re-evaluating CIB definitions in light of new AI technologies, and acknowledging AI's dual capacity to both create and detect CIB. Key enhancements include developing a comprehensive CIB detection toolkit, with the **Vera AI Alerts** tool specifically designed to identify coordinated social media networks through behavioural patterns and interaction metrics. vera.ai introduced a visual assessment report that applies 50 CIB indicators to quantify and visually represent CIB probability across five dimensions, and championed a shift towards adaptive, behaviour-focused detection methods, integrating advanced AI capabilities like LLM-based narrative extraction and multimodal similarity detection to counter evolving CIB tactics. Additionally, the framework references a number of external resources:
    - For reach and amplification analysis: MISP, BuzzSumo, Hootsuite, Social Blade, and OpenMeasures.
    - For content and sentiment analysis: NewsWhip, Lexalytics, and TextBlob.
    - For network analysis: Cytoscape, NodeXL, and Gephi.

[17] https://coornet.org
[18] https://cran.r-project.org/web/packages/CooRTweet/index.html
[19] https://coortweet.lab.atc.gr/auth/login
[20] https://www.disinfo.eu/wp-content/uploads/2024/08/20240805-CIB-detection-tree.pdf

- **Post-truth Space Mapping Technique Protocol**. Developed by the University of Amsterdam as part of the vera.ai project (Rogers & Koronska, 2025), this technique identifies and visualises clusters of problematic sources (dubbed "post-truth spaces") on platforms like Facebook. It begins with thematic keyword queries followed by a network analysis to map clusters of pages, groups or accounts. These clusters are then assessed for influence (via an 'influence metric') through a "betweenness centrality measure" applied to clusters, indicating how much post-truth content penetrates mainstream discourse. The technique is especially valuable for surfacing low-visibility amplifiers and fringe communities, including digital creators and coordinated entities, aiding fact-checkers in identifying high-impact content and potential disinformers. These techniques have been used by a series of fact-checking organisations (during the reported data sprints and Summer/Winter schools) and deemed useful in their endeavours to source leads and curate source sets to follow.
- **CIB Typology Development Technique**. The methodology developed by Richard Rogers and Nicola Righetti (2025) builds on the work of Giglietto and colleagues (2020; 2021; 2023) in the study of coordinated inauthentic behaviour (CIB), expanding its focus beyond deceptive influence operations. This approach enables the analysis of a broader spectrum of actors, including media organisations, political activists, and advertising networks, to compare their highly coordinated activities and examine both similarities and distinctions in their behaviours (Rogers & Righetti, 2025). The authors introduce the concept of "manufactured attention" and identify various forms of CIB, such as coordinated media sharing, political meme dissemination, gambling promotions, and advertisement spamming via hijacked groups. Their typology development employs a mixed-methods approach, beginning with the identification of highly coordinated link-sharing networks. A qualitative investigation is used to define and categorise the actors involved, which is then paired with a quantitative analysis of the degree and structure of coordination.

  Applying this methodology to 19 Facebook communities, the researchers developed six distinct typologies: media groups, advertising networks, large public groups repurposed for ad-sharing, political supporters or critics, and an influence operation characterised by anonymous pages disseminating content from a disinformation source. The quantitative analysis revealed varying degrees of coordination, from fully centralised sharing to broader, more distributed participation, which helped distinguish tightly controlled networks from loosely organised ones.

  The findings show that different types of entities can exhibit similar coordination signatures, blurring the lines between legitimate and deceptive behaviour. A continuum of inauthenticity is observed, ranging from anonymous, undisclosed coordination to openly identified, grassroots-style participation. While preliminary and specific to a given time frame and platform, this typology provides a foundational framework for differentiating types of coordination and critically examining the boundaries of inauthentic behaviour.

Table 6 outlines how these tools and methodologies can assist in measuring the impact indicators described in Table 4.

Table 6 Match-making between tools and impact indicators

FRAMEWORK / TOOLS	PLATFORM-RELATED INDICATORS			CONTENT-RELATED INDICATORS		MEDIA	NUMBER OF COMMUNITIES	TYPES OF HUMAN AMPLIFIERS		BELIEF CHANGES INDICATORS		BEHAVIOUR CHANGES- OFFLINE ACTIONS (policy responses/calls to violence)	
	Number of platforms	Reach (views)	Engagement (shares, reactions)	Language	Format	Fringe or mainstream media		Public figures	Recurrent disinformers	Sentiment analysis indicating polarisation, institutional discredit	Limitation/ suppression of freedom of thought and speech	Public health or safety risk	Limitation/ suppression of freedom of thought and speech
DBKF	X			X	X	X							
Verification Plugin: Image Reverse Search	X												
Vera Alerts on Facebook*		X	X										
Verification Plugin: SNA tools (X and Meta)		X	X					X					
Verification Plugin: XNetwork	X												

CIB Detection Tree**	X	X	X			X	X		
CooRnet	X						X		
Verification Plugin: Multilingual Persuasion Technique Classifier				X				X	X
Verification Plugin: Multilingual News Article Framing Classifier***				X					

** This tool has limitations due to CrowdTangle shutdown and shift to Meta Content Library*

*** Based on tools developed both within and outside the project*

****Provides information about "frames" and topics covered in the text, which can refer to social indicators*

3 A Meta-analysis on Algorithmic Amplification on TikTok

3.1 Framing the Issue: Algorithms, Amplification, and Systemic Risks

3.1.1 Defining Algorithmic Amplification: from Inputs to Influence

The concept of algorithm is well-defined in technical terms: it is a "precise step-by-step plan[21]" or "a finite set of unambiguous instructions[22]" that takes input(s), performs a sequence of operations, and produces output(s). These definitions suggest a clear, deterministic logic – a cause-and-effect structure typical of "a computational procedure[23]". Ironically, such linearity becomes obscured when algorithms are deployed within the opaque environments of digital platforms (social media, search engines, and others). In these contexts, algorithms determine what content users see, in what order, and why – yet the underlying mechanisms are often inaccessible to anyone outside the company's black box.

What we often refer to in these settings is not simply the algorithm, **but algorithmic amplification: i.e., the process by which algorithmic systems selectively elevate the visibility, reach, or spread of certain content**. Except for some proprietary data (for instance, Frances Haugen's whistleblowing[24] on Facebook in Hao's 2021 piece), in most cases we can only infer the otherwise largely hidden criteria that guide amplification, including engagement, virality, preference, exposures, and more.

Crucially, **amplification is not per se a type of algorithm, but rather a consequence or function of how an algorithm operates**. However, unlike algorithms, the notion of amplification lacks a standardised or shared definition. Each platform implements its own logic for curating and prioritising content, with an evolving 'recipe' whose ingredients remain largely unknown.

3.1.2 Algorithmic Amplification as a Systemic Risk: Disinformation and Engagement Loops

Algorithmic amplification is a critical concept in the study of disinformation because it helps explain how and why certain false or misleading narratives gain disproportionate reach and impact online. Numerous studies show that triggering, polarising, and often inaccurate content spreads farther and faster on platforms, as threats and negative stimuli evoke stronger reactions, which algorithms are designed to reward. Moreover, **malign actors often exploit algorithmic amplification to increase the distribution of content (for example, by orchestrating Coordinated Inauthentic Behaviour[25])**, strategically exploiting platform dynamics to their advantage.

Similarly, recommender systems often prioritise inflammatory content that generates higher engagement. While disinformation can originate from various sources, it is often these engagement-

[21] https://en.wiktionary.org/wiki/algorithm
[22] https://www.ahdictionary.com/word/search.html?q=algorithm
[23] https://en.wiktionary.org/wiki/algorithm
[24] https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/
[25] https://www.veraai.eu/posts/report-revisit-coordinated-inauthentic-behaviour-detection-tree

optimised algorithmic systems that propel these narratives into mainstream visibility, shifting the debate from free speech to free reach.

Under the EU's Digital Services Act[26] (DSA), Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) are required to **assess and mitigate "the systemic risks stemming from the design, functioning, and use of their services**" (Recital 79 of the Digital Services Act). Hence, **algorithmic amplification can itself constitute a systemic risk,** while also exacerbating others – such as the spread of illegal content and negative impacts on fundamental risks and democratic processes. Understanding this mechanism is essential to ensuring greater accountability in platform design.

3.1.3 Why TikTok: The "For You" Page and Platform Design as a Risk Vector

Each platform operates with its own algorithmic architecture, shaped by business models, user behaviour data, and specific engagement goals. For instance, YouTube pioneered algorithms optimised for watch time (Thompson, 2020), encouraging binge consumption through endless autoplay and increasingly extreme recommendations. Google prompted the introduction of the notion of the "filter bubble[27]", personalising search results based on past behaviour, which can limit the diversity of information users are exposed to. Facebook was not the first to use algorithms, but it has mainstreamed the idea of the algorithmically curated feed since 2006 (D'Onfro, 2016). Moreover, Meta made one step forward with Instagram, whose various features (i.e., Feed, Stories, Explore, etc.) rely on different algorithms (Mosseri, 2023). However, TikTok has arguably set the new standard for algorithmic engagement. Its "For You" page, now widely imitated by X, Meta and others, is not only highly effective in capturing user attention but also central to what makes the platform so 'sticky', keeping users scrolling far longer than intended.

The "For You" Page algorithm appears to offer hyper-personalisation, learning from even the smallest signals like watch duration, replays, or pauses. What makes TikTok's algorithm particularly unique – and widely regarded as the most advanced in the industry – is its ability to rapidly adapt to user preferences with minimal input, delivering highly engaging content even before a user has followed anyone or indicated explicit interests. Its cold-start efficiency and unparalleled engagement design have set a new benchmark for platform recommendation systems, prompting imitation by rivals (e.g., Instagram's Reels and YouTube's Shorts). Moreover, TikTok's dominance among younger users increases its cultural and political impact, as the algorithm shapes the information diet and worldview of a generation (Kemp, 2024). The platform does not just reflect user preferences; it actively curates and constructs them, making its recommendation logic both powerful and, in many ways, obscure. Accordingly, this report focused on TikTok as a case study to explore the dynamics of algorithmic amplification.

[26] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065
[27] https://thedashdouble.com/stop-google-search-personalization-filter-bubble/

3.2 Empirical Insights: How Recent Research Understands Amplification

3.2.1 Method: Selecting Relevant TikTok Studies on Amplification

To support the analysis of TikTok's algorithmic dynamics, a focused literature review was conducted using arXiv, an open-access archive for scholarly publications. The search targeted the 50 most recent articles (published as of May 22, 2025), given that (1) they explicitly mentioned "TikTok" in the title, and (2) they address themes related to algorithmic amplification. As a result, six papers were selected:

1. Annabell, T., et al. (2025). *TikTok Search Recommendations: Governance and Research Challenges*. (arXiv:2505.08385v1). arXiv. https://doi.org/10.48550/arXiv.2505.08385.
2. Bauman, F., et al. (2025). *Dynamics of Algorithmic Content Amplification on TikTok* (arXiv:2503.20231v1). arXiv. https://doi.org/10.48550/arXiv.2503.20231.
3. Ibrahim, H., et al. (2025). *TikTok's recommendations skewed towards Republican content during the 2024 U.S. presidential race* (arXiv:2501.17831). arXiv. https://doi.org/10.48550/arXiv.2501.17831.
4. Masood, M., et al. (2025). *Counting How the Seconds Count: Understanding Algorithm-User Interplay in TikTok via ML-driven Analysis of Video Content*. (arXiv:2503.20030v1). arXIv. https://doi.org/10.48550/arXiv.2503.20030.
5. Mosnar, M., et al. (2025). *Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings*. (arXiv:2504.18140v1). arXiv. https://doi.org/10.48550/arXiv.2504.18140.
6. Vera, J.A. & Ghosh, S. (2025). "They've Over-Emphasized That One Search": Controlling Unwanted Content on TikTok's For You Page. (arXiv:2504.13895v1). https://doi.org/10.48550/arXiv.2504.13895.

3.2.2 Findings: Definitions, Dynamics, and Measurement Approaches

The goal of comparative overview of the six case studies is to assess whether and how each study defines algorithms and algorithmic amplification, and to summarise their key findings. Table 7 outlines each study's reference to algorithms and amplification, their conceptual understanding of how amplification operates on TikTok – particularly via its "For You" Page (FYP) – and the empirical insights they contribute to this growing area of research.

Table 7 Summary of the case studies

Study	Reference to algorithms	Reference to amplification	Understanding of algorithmic amplification	Findings
Annabell et al.	Search algorithms	Not explicitly mentioned but reference to algorithmic curation (and acknowledgement of moderation)	Functioning of search recommendations (i.e., pre-formulated search terms) in the homepage search bar and as an overlay on some videos	• Challenges to report/moderate search recommendations • Relevance and engagement affect search recommendations • Users lack agency over search recommendations and cannot disable them
Bauman et al.	TikTok as an "extreme case of an algorithm-driven ecosystem"	Content amplification	Increase of interest-aligned content encountered by a user over time on TikTok's FYP	• Strong amplification of interest-aligned content (i.e., personalisation) • Rapid personalisation within the first 200 videos or 1,5 hours of browsing • Content amplification varies across interests (influencing factors hypothesised: engagement, supply, type) • As interest-alignment increases, popularity decreases • As interest-alignment increases, diversity decreases
Ibrahim et al.	Recommendation algorithm	Not explicitly mentioned but reference to recommendation algorithms	Type of algorithmic recommended content users are presented with on their TikTok's FYP	• Likes, Comments, and Shares provide users with indirect agency over what they see • Ideological skew towards Republican-aligned content • Engagement metrics indicate a Democratic bias • Republican-conditioned accounts receive more ideologically-aligned content than Democratic-conditioned accounts • Democratic-conditioned accounts encounter more opposite-party content than Republican-conditioned accounts

				• Most recommendations criticise the opposing party rather than promoting one's own
Masood et al.	Recommendation algorithm	Not explicitly mentioned but reference to personalised recommender algorithms	Temporal evolution of algorithm-based recommendations and their interplay with user experience on TikTok's FYP	• As personalisation (based on engagement) increases, popularity decreases • Likes, comments, or shares (explicit feedback) do not change future recommendations over (short or long) time • Recent views (implicit feedback) influence a user's inclination to like or share a video (explicit feedback) • Users like content changes • Breaking the recommendation sequence degrades user experience • Video content analysis can predict if a user will watch a video with 70% accuracy
Mosnar et al.	Algorithmically-curated recommendations	Not explicitly mentioned but reference to content personalisation	Algorithmically-curated content on TikTok's FYP based on implicit or explicit user feedback	• Strong to moderate effect of location on personalisation • Watch (implicit feedback) as a stronger personalisation factor than follow and like (explicit feedback) • Recommendation as exploration (i.e., diversity) at the start, followed by exploitation (i.e., personalisation) after interacting with around 1000 videos
Vera & Ghosh	Algorithmic recommendation systems	Not explicitly mentioned but reference to hyper-personalised recommendation of interest-based content	Algorithmic persistence as difficulty in removing unwanted content from TikTok's FYP	(All findings are based on user perception) • The recommender system over-amplifies a one-time search • The recommender system amplifies popular content • The recommender system amplifies unwanted content by extension (e.g., users align with the opposite view) • The recommender system amplifies content from the user's network • The recommender system amplifies content that was accidentally rewatched

In these studies, algorithmic amplification is generally understood as the recommendation of content or searches based on a combination of user signals. Therefore, Table 8 summarises the key factors identified across the case studies as influencing algorithmic recommendations on TikTok.

This synthesis reflects the elements most explicitly addressed or prioritised by the authors in each study. However, it is important to note that this does not imply other factors are absent or irrelevant – only that the focus of the analyses lay elsewhere. Given the black-box nature of TikTok's algorithmic systems, researchers are only able to capture partial insights through observation and inference, rather than a complete picture of how recommendations are generated.

Table 8 Factors that influence algorithmic recommendations on TikTok, according to the case studies

	User interests	**Content popularity**	**Explicit feedback (likes, shares, comments)**	**Implicit feedback (watch)**	**Time spent on the platform**
Annabell et al.		X*	X**		
Bauman et al.	X	[X]***			X
Ibrahim et al.	X		[X]****		
Masood et al.	X	[X]***		X	
Mosnar et al.	X		[X]****	X	X
Vera & Ghosh		X		X	

** The paper refers to "relevance", which we assume to mean popularity.*
*** The paper refers to "engagement" with search recommendations, which we assume to consist of explicit feedback.*
**** It decreases over time.*
***** It has an indirect or secondary effect.*

3.2.3 Discussion: Common Patterns, Gaps, and Challenges

Across the six research articles analysed, there is clear consensus around the concept of **algorithmic amplification**, even if the term is not always explicitly used. Most studies understand algorithmic amplification as the process by which content is selectively elevated, in terms of visibility, reach, or repetition, based on implicit or explicit signals. Whether described as personalisation, recommendation, or content curation, the studies uniformly identify TikTok's algorithm-led "For You" Page as a central mechanism for amplification. It is worth noting that the study on search recommendations (Annabell et al., 2025) highlights challenges in reporting and content moderation mechanisms, indirectly suggesting that a**mplification can also be understood in terms of de-amplification (e.g., downranking, shadow banning, recommendation ineligibility),** although this is not the main focus of the present report or the literature reviewed.

Moreover, even though only a few studies (Annabell et al., 2025; Bauman et al., 2025; Mosnar et al., 2025) make an explicit reference to artificial intelligence, there is a collective understanding that algorithmic amplification is fundamentally a product of AI-driven systems.

The idea that users are presented with **interest-aligned content** is widely accepted across the case studies (Bauman et al., 2025; Ibrahim et al., 2025; Masood et al., 2025; Mosnar et al., 2025). TikTok's algorithm demonstrates a remarkable capacity to learn user preferences after only a short period of use, often just a few hours of scrolling (Bauman et al., 2025; Mosnar et al., 2025). A recurrent finding is that, over time, users are funnelled into increasingly niche content that reflects their inferred interests rather than content popularity, which is nonetheless the only platform-driven factor among various user-driven elements. This leads to **reduced exposure to diverse topics** and the potential formation of echo chambers, reinforcing narrow patterns of consumption (Bauman et al., 2025; Masood et al., 2025).

Interestingly, the studies also converge on the **limited influence of active user engagement** – such as likes, shares, and comment – on the algorithm's recommendation outcomes (Ibrahim et al., 2025; Mosnar et al., 2025). Instead, the platform appears to rely more heavily on implicit feedback, particularly the act of watching a video (Masood et al., 2025; Mosnar et al., 2025; Vera & Ghosh, 2025). While the precise weighting of these signals remains unknown, one outlier observed is the algorithm's **tendency to over-amplify unwanted content** despite active user resistance (Vera & Ghosh, 2025). On the one hand, this disconnect between user intention and algorithmic outcomes offers an alternative view to the picture of successful hyper-personalisation most studies paint. On the other hand, it also confirms the limited power of explicit feedback mechanisms and, therefore, raises broader questions about the extent to which users can meaningfully influence the recommendations they receive.

These concerns point to a larger and recurring issue across the studies: the ambiguity of TikTok's algorithmic architecture and the challenges it poses to researchers. The evidence – generated by small-scale sockpuppet bot simulations (Bauman et al., 2025; Ibrahim et el., 2025; Masood et al., Mosnar et al., 2025) or by qualitative interviews that convey folk theories (Vera & Ghosh, 2025) – is hindered by its limited scope and the inability to account for the full variability of user experiences. Furthermore, the dynamic nature of TikTok's algorithm, with frequent interface updates and new feature rollouts (such as search recommendations), **undermines the reproducibility and generalisability of research findings over time**.

Collectively, these limitations point to a **structural gap: the lack of transparency and user control** over how TikTok generates, curates, and moderates recommendations – especially in emerging functions as search recommendations. This minimal agency and the decreasing diversification of content (otherwise appreciated by users) pose **risks around the amplification of harmful, sensationalist, or polarising content, including mis- and disinformation**. This may be strategically exploited, as seen in cases where the algorithm has been used to elevate political ideologies or enable coordinated inauthentic behaviour.

In sum, while the body of research analysed affirms a shared understanding of algorithmic amplification as a process that reinforces engagement through inferred user preferences, it simultaneously highlights the **urgent need for more transparent, auditable, and participatory models of platform governance**. Without greater access to the inner workings of recommendation (and moderation) systems, **efforts to assess and mitigate systemic harms or ensure accountability** will continue to be hindered by methodological uncertainty and a reliance on speculation.

3.2.4 Accountability in Practice: What TikTok's DSA Reports Reveal (and omit)

Half of the selected studies reference the Digital Services Act (DSA), particularly in relation to the need for greater transparency in recommender systems (Annabell et al., 2025); concerns around the platform's failure to prevent disinformation and coordinated inauthentic behaviour (Ibrahim et al., 2025), and the obligation to audit algorithmic systems that may pose systemic risks (Mosnar et al., 2025).

In this context, we examine **how TikTok's 2023[28] and 2024[29] Risk Assessment Reports address systemic risks linked to algorithmic amplification**. Both reports were produced in compliance with Arts. 34, 35, and 42 of the DSA and were criticised[30] for their vagueness and lack of meaningful insights. Despite their extensive length, they lack charts, figures, or data visualisations and the little numerical information available is presented in purely descriptive terms. Moreover, they contain redacted sections that limit transparency further.

Both reports include a standardised description of the FYP – named "The For You Feed ('FYF')" – as a personalised recommendation system influenced by factors such as user interactions (likes, shares, comments), searches, content diversity, and video popularity. However, while algorithmic systems are recognised as a source of potential systemic risk, **the analysis of these risks remains primarily content-focused** (i.e., what is being amplified) – rather than examining the design of the recommendation architecture itself (i.e., how it is being amplified).

The reports address a wide range of content-based risk modules, focusing on the protection of minors, elections, fundamental rights, and intellectual property, as well as the avoidance of gender-based violence content, terrorist content, illegal hate speech, and misinformation.

Yet, mitigation measures related to algorithmic amplification – where present – are largely declarative, weakly evidenced, and often repetitive. For example, the reports mention the FYF's ineligibility rules (e.g., for content created by underage users, or misinformation awaiting fact-checking), but provide little insight into how these rules are technically implemented or audited.

Efforts to adapt algorithmic systems, including recommender systems, and thus mitigate algorithmic risks comprise the ability to reset the FYP/FYF experience and features such as the "Why this video?" explanation tool, or the use of an "unverified content" label. Labelling renders content ineligible to appear on the FYP/FYF, while still allowing it to be found via search – though it remains unclear whether this applies to search recommendations as well. A manual review is applied to content surpassing a certain popularity threshold, and users can opt to exclude content based on specific keywords (e.g., in cases of gender-based violence or hate speech).

The 2024 report introduces minor updates over its predecessor, including the identification of **AI-generated content and recommender systems as critical risk drivers**. Nevertheless, it asserts that the platform has adopted "a diligent and cautious approach to the analysis of systemic risks that may arise from the design, functioning, use or mis-use" of the service. For instance, TikTok claims to have adapted its algorithmic systems and implemented an additional automated content moderation model to detect hate speech, which is ineligible for recommendation in the FYF/FYP – although it remains accessible via search.

TikTok also recognises **"concentrated content" as a relevant risk** embedded in recommender system design, noting its potential to reinforce negative personal experiences (with references to body image, dieting, and fitness). However, the platform emphasises the difficulty of identifying and addressing

[28] https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwlY_fjulyhwzuhy%5B/ljhwZthlaukjlkulzlp/DSA_H2_2024/TikTok-DSA-Risk-Assessment-Report-2023.pdf
[29] https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwlY_fjulyhwzuhy%5B/ljhwZthlaukjlkulzlp/DSA/TikTok_DSA_Risk_Assessment_Report_2024.pdf
[30] https://dsa-observatory.eu/2024/12/09/dsa-risk-assessment-reports-are-in-a-guide-to-the-first-rollout-and-whats-next/

undiversified content, citing its non-violative nature according to guidelines and potential implications for freedom of expression.

Ultimately, these reports fall short of clarifying how TikTok's recommender systems operate in practice. The vague references to "user interests" and "historical engagement" – already inferred through observation and experimental studies – offer little concrete evidence of risk mitigation at the design level.

In sum, while the platform acknowledges algorithmic systems as vectors of risk, its disclosures remain superficial, anecdotal, and heavily content-centred. The systemic risks surfaced by the case studies are largely overlooked, except through a permissive framing that neither quantifies the problem nor treats it as a design-related concern, but rather as an inevitable and acceptable platform feature.

In reality, these reports offer no new insight into how content curation and recommendation on the FYP function. Without meaningful transparency around algorithmic design or empirical substantiation of mitigation measures, **these DSA risk assessment reports do not provide adequate visibility into how algorithmic amplification is being managed**. This raises broader concerns about enforcement gaps under the DSA, where formal compliance through report submission masks substantive opacity in algorithmic design and mitigation.

4 TikTok algorithmic amplification Case studies

4.1 Algorithmic Audits of TikTok: On Poor Reproducibility and Short-term Validity of Findings

This case study examines the reliability of algorithmic audits on TikTok's recommendation system, focusing on the reproducibility and generalisability of their results. As platforms increasingly tailor content through algorithms, researchers and policymakers have emphasised the need for systematic audits. However, our research shows that attempts to replicate earlier audits face obstacles from both evolving platform behaviour and limitations in prior research methods. Moreover, the findings from such audits often prove short-lived, highlighting the need for adaptable and consistently reproducible auditing approaches to track changes over time. Among the state-of-the-art TikTok studies reviewed in the meta-analysis presented in the previous section, the following section highlights the contribution of Mosnat et al. (2025), developed within the framework of the vera.ai project.

4.1.1 Rationale

The research responds to growing concerns over the opacity and influence of algorithmically curated content on platforms such as TikTok, especially given their shift toward short-form video and implicit user feedback. Given how platforms increasingly shape user experiences, regulators (notably through the EU's Digital Services Act) and researchers are calling for systematic algorithmic audits to assess how recommendation systems work and evolve.

The central rationale behind this research is to evaluate the **reproducibility** and **generalisability** of findings from existing TikTok algorithmic audits, particularly in light of platform evolution, policy changes, and regional variation. By replicating and extending prior studies on the platform's For You recommendation feed, we aimed to determine how consistent audit results remain over time and across settings. This involves adapting to shifting platform features and refining previous methodological approaches.

The study is guided by two core research questions (RQs) (Mosnar et al., 2025, p. 2):

1. *What is the level of reproducibility of algorithmic audits taking into consideration the constantly evolving nature of social media platforms?*
2. *What is the effect of different personalisation factors, such as location, watch duration, liking and following, on the For You video recommendation on TikTok? How has the importance of different personalisation factors evolved since the reference studies?*

To address these research questions, we replicated previous TikTok audit studies but introduced several necessary extensions and adjustments. These included expanding the study across multiple countries to account for regional algorithm differences, correcting methodological flaws such as video watch time inconsistencies and shadow-banning issues, adapting the audit tools to recent platform changes (like ad and livestream handling), and conducting additional data analysis, including a post-audit review using bot

data accessed via GDPR requests, to gain deeper insight into platform behaviour (Mosnar et al., 2025, p. 2).

We argue that reproducibility is essential for audits to be actionable and policy-relevant. Without it, regulators cannot track or verify the effects of interventions or platform changes. Our findings expose major challenges in replicating past audits, leading us to call for a shift toward longitudinal, reproducible, and automated audit practices. Essentially, the study advocates for "reproducible, longitudinal, multiplatform and more authentic (in terms of user simulation) audits" (Mosnar et al., 2025, p. 2) as an approach to distinguishing whether shifts in audit outcomes stem from changes in content, platform policies, or algorithmic behaviour.

4.1.2 Methodology

While attempting to replicate the methodology and code from earlier TikTok audit studies based on agent-based sockpuppeting algorithmic audit of the personalisation factors on TikTok, we encountered significant reproducibility issues, some of which obstructed the audit process entirely. To address these obstacles, we redesigned the study's methodology by updating the audit code to reflect platform changes, refining how personalisation factors were tested, and modernising the simulated user interests. We also enhanced the data analysis approach. These methodological improvements enabled us to identify what limited reproducibility (addressing RQ1) and to effectively compare personalisation effects between their study and prior audits (addressing RQ2) (Mosnar et al., 2025, p. 3).

To investigate the role of different personalisation factors on TikTok, we carried out an **agent-based sockpuppeting audit,** where automated user agents (bots) engaged with TikTok's *For You* page via the web interface. Each bot operated under a newly created user profile with randomly assigned personal details, including name, email, birth date, and a specific location. To assess the role of different personalisation factors, we defined a series of audit scenarios that altered the general behaviour or characteristics of the bots (see Table 9). These scenarios were largely based on a prior reference study, with redundant or overlapping cases intentionally removed to enhance reproducibility. The selected scenarios corresponded to the following personalisation factors: none (to control for noise), location, watch duration, liking and following (Mosnar et al., 2025, p. 3).

In each scenario, two bots were run in parallel:

- A control bot, which remained neutral.
- A personalised bot, which interacted according to the designated factor.

Each scenario involved four sessions (runs) per bot, with the same user account logging in for each run. Sessions were spaced approximately one day apart to give TikTok's recommender system time to adapt and personalise content. User history was preserved either by reusing saved cookies, or logging in again via the platform's interface. Data collection was conducted during January and February 2025. To confirm the consistency of results, some scenarios were repeated with a different pair of control and personalised bots. These replications produced no significant variation, confirming the robustness and reliability of the audit methodology.

Table 9 Scenarios for investigating the effects of personalisation factors, with information on how long the bot watches the videos and how many times the scenario is repeated (column Rep.) (Mosnar et al., 2025, p. 5). Note that the "control" specifies both the watchtime of the control user as well as the videos deemed "not relevant" for the personalised user. [S#] denotes ID of each scenario.

Factor	Scenario setup	ID	Watchtime Control | Personalised	Rep.
None	Country: USA	[S0]	100% | N/A	1
Location	Country: Germany	[S1]	100% | 100%	1
	Country: France	[S2]	100% | 100%	1
	Country: Romania	[S3]	100% | 100%	1
	Country: Ukraine	[S4]	100% | 100%	1
Watch	25 random videos	[S5]	25% | 50%	1
	25 random videos	[S6]	25% | 75%	1
	25 random videos	[S7]	25% | 100%	1
	25 random videos	[S8]	100% | 200%	1
	Hashtag: movie, film, marvel, foodtiktok, tiktokfood, foodie, cooking, food, gaming, gta6, gta,	[S9]	25% | 50%	1
	minecraft, roblox, cat, dog, pet, dogsoftiktok, catsoftiktok, cute, puppy, dogs, cats, animals,	[S10]	100% | 200%	2
	petsoftiktok, kitten, comedy, lol, humour, laugh, fun, jokes, love, couple, relationships	[S11]	100% | 400%	2
Like	10 random videos	[S12]	100% | 100%	1
	Hashtag: Same as watch	[S13]	100% | 100%	2
	Creator: Top 30 creators on TikTok	[S14]	100% | 100%	2
Follow	5 random creators	[S15]	100% | 100%	1
	10 random creators	[S16]	100% | 100%	1

4.1.3 Main Findings

RQ1: Reproducibility of Algorithmic Audits

Our study identifies two major groups of factors that negatively affect the reproducibility of algorithmic audits (Mosnar et al., 2025, pp. 5-6):

Audit-specific factors:

- The most critical issue is the **availability and quality of resources**, including the **audit code** and the **videos** encountered during previous audits. Frequently, these are either not publicly released or incomplete.
- The **published source code** often misses fundamental parts, requiring **reverse-engineering** (e.g., reconstructing missing database schemas).
- **Incomplete or inaccessible repositories** further prevent effective reproduction.
- A lack of detail in **study design and methodology**, such as missing specifics on **watch duration** or **user characteristics**, limits the precision of replication.
- Inconsistencies between **code and paper descriptions** create ambiguity.
- Methodological flaws in the original studies (e.g., stronger feedback on non-interest content) also compromise reproducibility.

To address these factors negatively affecting audits' reproducibility, we recommend adhering to **reproducibility practices** from the **machine learning** field.

Platform-specific factors:

- **Content evolution** on platforms like TikTok leads to rapid changes in what is recommended, making audits conducted even days apart yield different outcomes.
- **External events** (e.g., elections) can further bias content, requiring careful control of content diversity.
- **Platform evolution** introduces features like livestreams and ads, altering bot experience. These elements appeared more frequently in this audit (e.g., 1 in 4 videos were ads), complicating comparisons with previous studies.
- **Changes in HTML structure** (including country-specific variations) frequently break audit implementations.
- Increasing **bot detection and banning** measures were observed. In several cases, bots were banned mid-audit, with simulated users located in **Italy** being particularly problematic. Post-audit checks showed additional bans (especially for bots with implicit-only actions), partly due to proxy server issues.

To address these factors, we stress the importance of **post-audit integrity checks**, including GDPR data requests, which revealed that explicit actions (likes, follows) were often missing from user history when conducted via the **web interface**. This impacts the audit's validity, as these actions may not influence the recommender system as expected.

RQ2: Effects of Personalisation Factors

- **Controlling for noise**. The default scenario (S0), where both control and personalised users behaved identically, was used to establish a baseline. This setup allowed us to measure the inherent variability or "noise" in recommendations by comparing overlap between two identical user accounts. Approximately 30% of videos overlapped, consistent with prior work, confirming the platform introduces variability even without behavioural differentiation.
- **Personalisation factor: location**. Location had a moderate personalisation effect, strongest at the beginning of sessions when the recommender system relied on initial metadata (like IP). Some countries (e.g., Romania, Ukraine) showed more topical divergence, possibly due to events like elections or war, but effects diminished over time as behavioural feedback took over.
- **Comparison of implicit and explicit personalisation factors**. The study clearly distinguishes between implicit feedback (e.g., watch duration) and explicit feedback (e.g., likes, follows). Watch duration, an implicit factor, had the strongest and most consistent personalisation effect. In contrast, likes and follows showed minimal or no effect, likely due to limitations in how TikTok registers explicit actions on the web interface. This contrast highlights the limited impact of explicit personalisation in this setup, deviating from earlier findings and underscoring the importance of using mobile-based audits for evaluating such actions.
- **Ratio of videos of interest**. In interest-based scenarios, bots watched or interacted more with videos that matched predefined hashtag interests. Over time, personalised users received more "videos of interest", while control users maintained a lower, consistent ratio. This confirms the recommender's ability to align content with user interests, particularly in watch-duration-based personalisation.
- **Effect of different watch duration**. A set of scenarios tested the impact of varying watch times:
    - Shorter watches (25%, 50%, 75%)
    - Full watches (100%)
    - Rewatching (200%, 400%)

  Longer watch times led to stronger personalisation, seen through lower average video popularity (i.e., more niche content) and higher hashtag similarity. Particularly in interest-based setups, longer durations sharply increased content alignment. **Watch duration is the most influential personalisation signal**, especially when paired with interest-specific content (Mosnar et al., 2025, pp. 7-9).

4.1.4 Case Study Summary

In the process of replicating the algorithmic audit to assess the consistency and broader applicability of its results, we identified two key challenges: 1. poor reproducibility and 2. short-term validity of findings (Mosnar et al., 2025, p. 9):

1. **Poor reproducibility**. We identified multiple factors that negatively affect the audit reproducibility, stemming both from the platforms themselves, but also from the audit methodologies. The constantly evolving platform content, presence in multiple countries with different policies on AI or recommender systems, evolving platforms due to the added functionality or due to the policy changes, or the active fight of the platforms against automated bots – all these factors contribute to the poor reproducibility. Even during the short time, when we were working on the audit reproducibility, we had to modify the underlying audit code to adapt to these changes. The poor reproducibility is further exacerbated by the limited resources released by the previous studies, limited description of their study design and methodology, or different methodological choices. As such, the algorithmic audits in the current state cannot be easily repeated and need to be constantly adjusted, which limits their usability and requires extensive work.
2. **Short-term validity of findings**. Reimplementing and rerunning the reference studies, we find a significant change in the overall findings. Compared to these studies, we have found that the impact of the watch action is significantly stronger, which is further increased as the video is watched for longer or multiple times. At the same time, we observe a shift to a stronger exploration component for the like and follow action, which reduces their personalisation impact in the first set of around 1000 videos. However, we also observe that some explicit actions may not be taken into consideration by the recommender algorithm when using TikTok's web interface, as these actions do not appear as part of the requested GDPR data. Most significantly, the findings are strongly dependent on the methodological setup, where just by changing the evaluation metric or how strict we are when computing similarity may lead to completely different findings. A similar effect can be observed based already with small variations in the bot simulation.

Thus, we "advocate for reproducible, longitudinal, multiplatform audits with more authentic user simulation that can more faithfully discern the changes in findings" (Mosnar et al., 2025, p. 9). The case study card is presented in Table 10.

Table 10 Algorithmic Audits of TikTok: On Poor Reproducibility and Short-term Validity of Findings case study card

Category	Element	*Detail options*
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☐ *High* ☑ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☑ *Yes* ☐ *No* ☐ *Not Investigated*
	Type of Amplification (if any)	☑ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: _____*
Data Collection	Method Used	☐ *API* ☐ *Scraping* ☐ *Third-party tool* ☐ *Manual* ☑ *Other: agent-based sockpuppeting audit*
	Platform / Tool	*TikTok*
	Data Collection Time Structure	☐ *Single snapshot* ☐ *Longitudinal* ☐ *Continuous* ☑ *Experiment-based*
	Data Source Selection	☑ *Expert-driven* ☐ *Hashtag/topic-based* ☐ *Automated search*
Data Analysis	Type of Analysis	☑ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☐ *Mixed methods*
	Depth of Analysis	☐ *Surface-level* ☐ *Thematic coding* ☐ *Behavioural network analysis* ☑ *Other: personalisation effects on recommender system*
Limitations & Open Questions	Identified Limitations	*platform restrictions, reproducibility, incomplete data, time constraints*
	Key Open Questions	*long-term effects, data access issues, recommender system opacity*

4.2 TikTok Comments War: Signs of Inauthentic Coordination in the Context of the Second Round of Presidential Elections in Romania?

4.2.1 Rationale

In the wake of a tumultuous cancellation of the first round of presidential elections, TikTok became a widely used – and contested – platform in the context of the reorganised election round. The online presence of previous presidential candidate Calin Georgescu, whose campaign was massively boosted on

TikTok and alluded to all the legal transparency requirements, ultimately led to the cancellation of the election results before the second round of voting took place. The far-right candidate was allegedly boosted within the online campaign with Russian support, according to declassified intelligence documents. As a result, Romania's Constitutional Court annulled the election results, citing concerns over foreign interference and electoral irregularities, despite heavier monitoring by Romania's Permanent Electoral Authority and Central Electoral Bureau during the second round of the first election. Taken together, due to the poor management of the political crisis and lack of transparency in decision-making, the situation escalated into a massive democratic breakdown and led to a new round of elections. As a result, TikTok became a key platform for investigation to investigate for signs of inauthentic coordination.

Additionally, the foreign influence and bot army allegations became common narratives in the presidential campaign and its coverage, with recurring and varied iterations. Therefore, this case study aims to explore how the comment sections from the posts of the most important five political candidates exhibit traces of coordination behaviour, and the extent to which they could be considered inauthentic. Do commenters comment on all or nearly all posts? This would be a strong indication of coordinated campaigning. Are TikTok commenters sharing copypasta or near duplicated content rapidly? Do their accounts have recent, near simultaneous creation dates? These would raise questions about the authenticity of the coordinated action.

4.2.2 Methodology

The aim of the data collection is to gather an extensive set of comments pertaining to the candidates in the second round of elections. To this end, we conducted periodic scraping of all posts dating from 15 March 2025 until the elections, as detailed below. Five candidates were selected for the study, according to these criteria: the opinion poll ranking and the TikTok account usage as an instrumental element of the campaign, as well as the engagement it attracted.

The start date chosen represented the filing deadline for candidacies in the presidential elections. The end date for the scraping varied depending on the candidate's chances of reaching the second round of elections, and the engagement rate of TikTok. Thus, for three out of five candidates, the end date was 29 April 2025, as a result of the opinion polls published in the 21-28/04 period. For the remaining two candidates, who reached the second round, the end date was the day of voting, 18 May 2025.

The data was collected using Zeeschuimer (Peeters, 2023), a browser extension designed for researchers to study social media content. We used the scraper because access to the TikTok API has proven problematic for researchers (Entrena-Serrano et al., 2025; Pearson et al., 2024). The browser extension collects data from the platform's interface as the user scrolls. This approach is especially useful for platforms that limit or block traditional scraping methods (or have limited API access). The data scraped with the Zeeschuimer extension was uploaded to the 4CAT Capture and Analysis Toolkit, a research tool for analysing and processing data from online social platforms (Peeters & Hagen, 2022).

Table 11 shows the total amount of posts for each candidate, and the comments collected out of the total comment number. Due to the fact that TikTok is not designed to enable data scraping through its design, and to limit any signs of automated data collection methods, all datasets had to be collected manually

scrolling as much as the interface would allow. This also means that not all comments could be collected from each post. When a post exceeds several thousand comments, the interface restricts the scrolling. Almost 75% of comments were scraped.

The case study investigates four key avenues to identify coordination and its authenticity:

1. the coordination of their commenting behaviour through rapidity and regularity;
2. the coordinated sharing of the same (copypasta) or similar content in their comments (near duplicate detection) by a set of commenters;
3. the size of the group of power commenters and whether the power commenters target particular candidates.

For the analysis of the coordinated sharing of the same or similar content in the TikTok comment sections, we utilised **LLM-assisted prompting** to detect repeated or near-duplicate comments. As such, a LLM was guided through a structured prompting sequence: first, to normalise comment text (lowercasing, punctuation removal), then to identify exact duplicates and near-duplicates using string matching (e.g., Levenshtein distance). Follow-up prompts directed the LLM to group similar comments, count how many users reused them, and assess whether these repetitions occurred across multiple posts or in tight time windows. This prompting-based method enabled the detection of repeated messaging, a core indicator of coordinated campaigning.

We analysed commenter activity, calculating the **volume and regularity of comments per user** and measuring engagement across posts and over time. “Power commenters” are defined here as users who posted at least 10 comments in the dataset. One pattern that was analysed was the commenting behaviour, specifically if users systematically commented on posts from specific candidates, focusing on frequency, breadth (number of posts targeted), and timing (delay between post and comment).

To check whether power commenters had particular targets, all comments from the top 500 power commenters were matched to post authors using post IDs. The data was then grouped by commenter and candidate to count how many unique power commenters engaged with each candidate. This allowed for the identification of concentrated patterns of activity, such as large clusters of power commenters exclusively engaging with a single candidate, which served as indicators of targeted, potentially orchestrated engagement. A **network graph** was constructed to visualise patterns of interaction between power commenters and candidate accounts (see Figure 2).

Limitations

- The scraping method used for the current case study is affected by the platform-based scrolling limit.
- Scraping was conducted on the desktop web version of TikTok, rather than the widely used mobile experience.
- Given the long scraping process inherent in such a method, there is the risk of data loss due to user action or platform content moderation such as the deletion of comments.

4.2.3 Main Findings

Table 11 below showcases the total number of posts and comments scraped relative to the total comment number. Figure 1 visualises the comments amassed by each post per candidate in the indicated timeframes. The data analysis presented below reveals multiple layers of potentially inauthentic coordination within the TikTok comment sections related to Romania's second round of presidential elections. From repeated message patterns to strategic targeting by hyperactive accounts, several digital behaviours suggest organised efforts to influence the discourse around specific candidates.

Table 11 Romanian candidate posts and comment number overview

Candidate	Date Range	Post number	Scraped comment number / Total Comment Number
Nicusor Dan	15/03/2025-18/05/2025	185	**94,527**/ 14,1264
George Simion	15/03/2025-18/05/2025	49	**137,712** / 16,1163
Crin Antonescu	15/03/2025-29/04/2025	107	**5,679** / 5,679
Victor Ponta	15/03/2025-29/04/2025	218	**61,022** / 89,001
Elena Lasconi	15/03/2025-29/04/2025	139	**33,955** / 54,399
Total	15/03/2025-29/04/2025	698	**332,895** / 451,506

Post and comments activity

The posting activity was marked by various posting patterns, as seen in Figure 1. George Simion, whose usual posting rate was considerably higher, had the lowest post number, while amassing the highest number of comments. It's worth noting that the posting activity of the key presidential candidates posed several **irregularities**: Nicusor Dan explicitly invoked cyber attacks and resorted to change the privacy of his account in the 15-17 April period. Additionally, George Simion also claimed that his account was taken down one day before the election, which was later on reported to have been closed by the user, not by the platform.

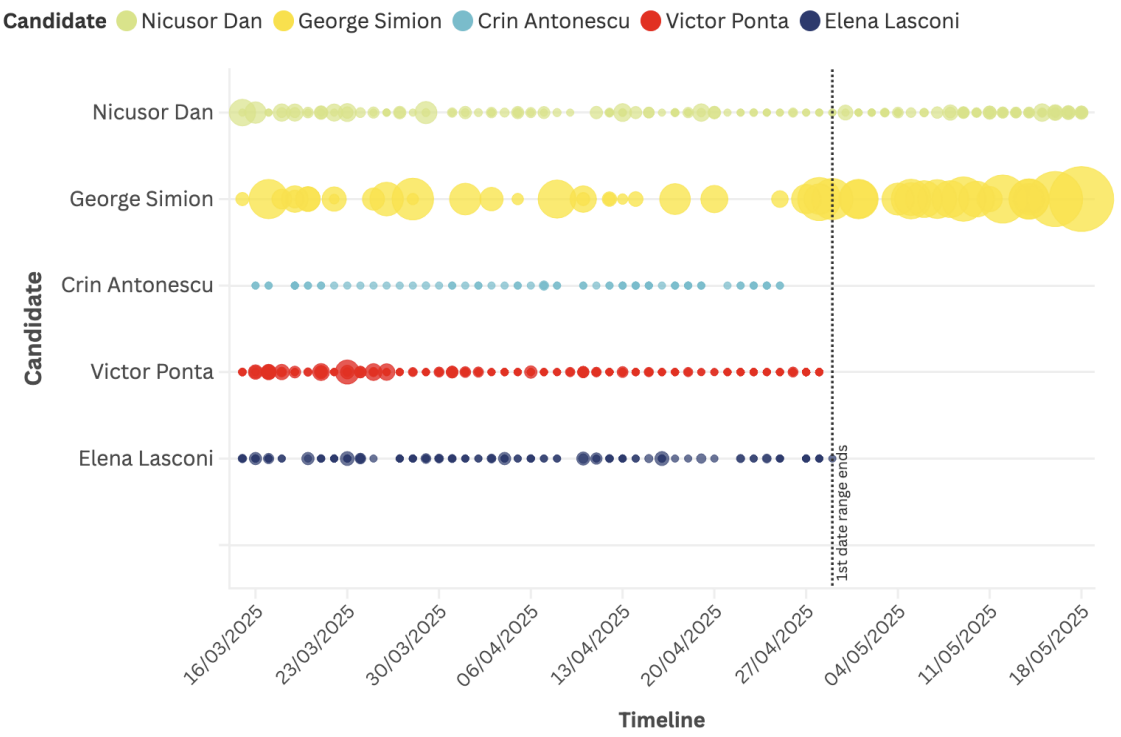


Figure 1 Post and comment activity plotting for key candidates; the dots represent the posts and the size represents the number of comments per post

Duplicate and near-duplicate comment detection

A first and clear indicator of coordination is the widespread use of identical or near-identical comments, often referred to as "copypasta". The analysis identified over **300 unique comment sets** that appeared repeatedly across multiple users. Top phrases include common slogans such as "[Candidate] president", "Bravo", "Respect", "Good luck" or voting intentions such as 'massively voting' and 'God bless'.; these comments typically express support or hostility toward a candidate using identical phrasing, punctuation, or emoji patterns. Additional similar comment phrases include the "round two back" phenomenon, a reverberation of the previous annulled election rounds.

Notably, the same message often appeared on different posts within minutes, suggesting possible use of automation or scripted coordination. Some of these comments were replicated by over **40 distinct users**, often targeting the same candidate or set of candidates. This pattern exceeds the expected behaviour of organic public discourse.

The **Coordinated Sharing Detection Service (CSDS)** was able to identify several clusters of identical comments, which can be parsed in two categories:

- **Sets of one to three identical emoji** usage, which can be authentic behaviour due to the fact that TikTok offers this type of comment nudges when a user attempts to leave a comment. One such example is represented by the chair emoji, alluding to George Simion's absence from the debates organised on the second round of elections.

- **Hieroglyphs with support messages**, as a means to limit algorithmic detection and thus bypass moderation policies.

Comment volume and the role of power commenters

Out of the 147,143 unique users in the dataset, **approximately 3,000** commenters posted more than 10 comments across all candidates' posts. These power commenters are responsible for **almost half of all comments**. The overview of the power user targeting per candidate is as follows:

- **Nicusor Dan** (nicusordanpmb) emerged as the most targeted by power commenters, amassing the highest number of unique power commenters, 1,236 out of the total 42,261 unique commenters, with **over 1,000** of them exclusively targeting him.
- **Victor Ponta** ranks second as the second most targeted candidatefollows, with 720 power commenters out of 28,721 unique ones.
- **George Simion** (georgesimionoficial) had a high number of commenters too, 81,360, with almost 700 power commenters. This shows more diverse but less repetitive support.
- **Elena Lasconi** (elena.lasconi) hit 425 power commenters out of the 15,658 unique ones.
- **Crin Antonescu** had a negligible number of comments amassed, without any power commenters.

There is **very little overlap** within the power commenters' target, with just 240 out of almost 3,000 targeting more than 1 and no more than 3 political candidates. Figure 2 provides a closer look at the comment target clustering among political candidates.

Suspicious power commenting patterns

- Out of the approximately 3,000 power commenters, almost **150** exhibited **high-frequency, tightly clustered comment bursts**, with 25 suspicious power users having a median time gap of **1-2 seconds between consecutive comments**, suggesting that their comments were made unrealistically close together. Such unnaturally short inter-comment gaps point to likely use of automation or copy-paste techniques.
- 25 suspicious power commenters were active for **1 day**, while the average active days for all the suspicious power users is **9 days**.
- On the other hand, despite long overall activity windows (spanning weeks), the accounts flagged as suspicious that were active for longer periods show **repeated micro-bursts**, indicating potential use of scheduled automation tools or bot-assisted commenting tactics.
- There are 13 suspicious power users with over 100 comments, whose commenting patterns suggest **constant or sustained output**, which is rarely consistent with organic user engagement.

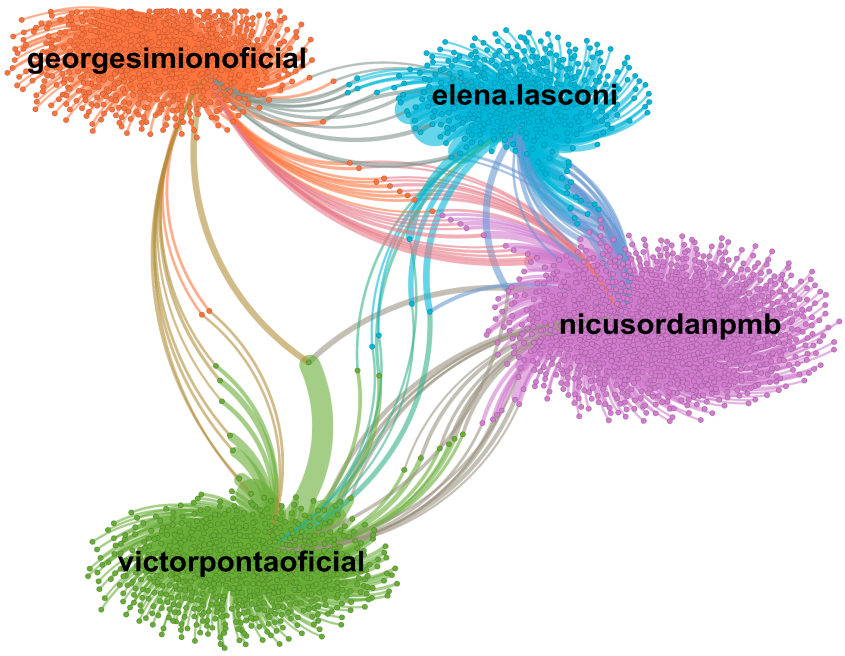


Figure 2 Network graph showing the clusters of power commenters targets across key political candidates, labelled over the four clusters; the nodes are highly active user accounts, and an edge connects a commenter to a candidate if they commented on that candidate's posts, and its thickness corresponds to the number of posts commented on

Conclusions:

- The analysis provides **empirical evidence of coordination, possibly inauthentic**, in the TikTok comment sections around the 2025 Romanian presidential elections, with **2% of users – power commenters – generating almost 50% of the total number of comments**. From repeated messaging and burst timing to user behaviour concentrating on specific political targets, the patterns observed go beyond typical online engagement. While not all activity is necessarily inorganic, the scale, structure, and synchronicity of many user behaviours strongly suggest centrally organised digital campaigning efforts.
- The technique helps fact-checkers and journalists quickly analyse election-related coordination. However, success depends on involving local experts with media literacy and language skills to accurately assess disinformation. These skills are key to distinguishing inauthentic behaviour from normal contextual factors like low media literacy, democratic engagement, and social media use patterns.
- **Access to data under DSA's article 40 remains limited,** making it difficult to monitor disinformation spread on TikTok in real time. The application procedures of the DSA Transparency Database[31] and of the TikTok Research API[32] involve long waiting time and rigid proposals, which may limit the scope of the data collection.

[31] https://transparency.dsa.ec.europa.eu/
[32] https://developers.tiktok.com/products/research-api/

Table 12 presents the case study summary.

Table 12 TikTok Comments War: Signs of Inauthentic Coordination in the Context of the Second Round of Presidential Elections in Romania? case study summary

Category	Element	***Detail options***
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☐ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☐ *Yes* ☐ *No* ☑ *Potential* ☐ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☑ *Engagement* ☑ *Other: commenting behaviour, power user activity*
Data Collection	Method Used	☐ *API* ☑ *Scraping* ☐ *Third-party tool* ☑ *Manual*
	Platform / Tool	*Zeesschuimer, 4CAT, CSDS, Gephi*
	Data Collection Time Structure	☐ *Single snapshot* ☑ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☑ *Expert-driven* ☐ *Hashtag/topic-based* ☐ *Automated search*
Data Analysis	Type of Analysis	☐ *Automated (keyword/topic clustering)* ☑ *Manual coding* ☑ *Mixed methods*
	Depth of Analysis	☑ *Surface-level* ☐ *Thematic coding* ☐ *Behavioural network analysis* ☐ *Other: _____*
Limitations & Open Questions	Identified Limitations	*platform restrictions, reproducibility, incomplete data, time constraints*
	Key Open Questions	*data access issues, ability to distinguish between inauthentic behaviour and low media literacy*

5 Inauthentic Coordination Impact Case studies

5.1 Influence by Design: How Coordinated Networks Spread Political Content on Polish TikTok

5.1.1 Rationale

This case study was developed by the University of Amsterdam in collaboration with Alliance4Europe for a Polish-government funded research agency to investigate indications of inauthentic behaviour in the online sphere. A coordinated campaign involving more than 2,000 accounts, over 600 of which showed synchronised activity, was detected on Polish X (formerly Twitter) just three months before the presidential election. The operation largely amplified right-wing and far-right TikTok videos.

Between 22 March and 6 May, the network produced over 30,000 posts linking to more than 1,400 TikTok accounts. Much of the content was politically charged, frequently critical of the government and Poland's geopolitical position. The strategy appeared "ambient" in nature, designed to flood the information space and create the illusion of broad public sentiment (Rogers & Righetti, 2025). The accounts behind the campaign bore multiple signs of inauthenticity: recent creation dates, low follower counts, and minimal original output.

At the core of the operation was cross-sharing: the reposting, repurposing, or adaptation of TikTok videos across X. TikTok's native share feature facilitates this process by allowing videos to be distributed within the app and externally to other platforms, such as: X, Instagram Stories, Reddit, and messaging platforms.

Such cross-platform sharing contributes to algorithmic amplification. Creators themselves often encourage off-platform redistribution to boost engagement, effectively transforming X into an informal syndication network. Empirical evidence shows that this practice increases content visibility, as TikTok's recommendation system prioritises videos based on viewing time, likes, comments, and shares (Klug et al., 2021).

By mobilising a large network of X accounts to circulate TikTok content, operators were able to artificially inflate these metrics, nudging TikTok's algorithm to further promote the videos. TikTok's share function also adds a layer of obfuscation: every share generates a unique short link, complicating detection of coordinated activity.

However, the process is not fully opaque. Automated shares append standardised captions, typically with phrases like "Watch the video", the original account handle, and the #TikTok hashtag. When paired with timestamps, URLs, and behavioural data, these markers enable researchers to identify large-scale cross-platform patterns.

It was through such digital traces that the present investigation revealed both the scope and the coordination of this network designed to amplify political messaging across TikTok and X.

5.1.2 Methodology

To assess whether the observed activity amounted to coordinated behaviour, we applied the CIB detection tool CoorTweet. The tool analysed a dataset of more than 30,000 posts on X (formerly Twitter) containing the standardised caption "Watch the video" („Zobacz filmik użytkownika"), collected between 22 March and 6 May. This phrase was selected because it commonly appeared when TikTok videos were automatically re-shared to X.

Following recent methodological advances, coordination was tested across two temporal windows – 20 minutes and 24 hours – in order to capture both short-term and longer-term synchronisation.

The tool's output generated clusters of X accounts along with associated metadata. These clusters were subsequently examined using basic quantitative techniques. Key account-level attributes, such as account creation dates, follower and following counts, were further enriched with supplementary data retrieved from Junkipedia[33] and Meltwater[34].

5.1.3 Main Findings

The investigation uncovered a coordinated network of more than 600 accounts on Polish X amplifying right-wing and far-right TikTok videos through the platform's auto-share feature. Activity varied: some accounts produced vast numbers of posts to amplify political messages, while others appeared dormant or automated, recycling content riddled with errors.

[33] A civic technology platform for monitoring and analysing digital content, designed to help researchers and civil society organizations identify, collect, and archive problematic material such as misinformation, hate speech, and junk news across social media, fringe networks, and messaging apps.

[34] A commercial media intelligence company offering social listening, analytics, and AI-driven insights across media, social, consumer, and sales domains.

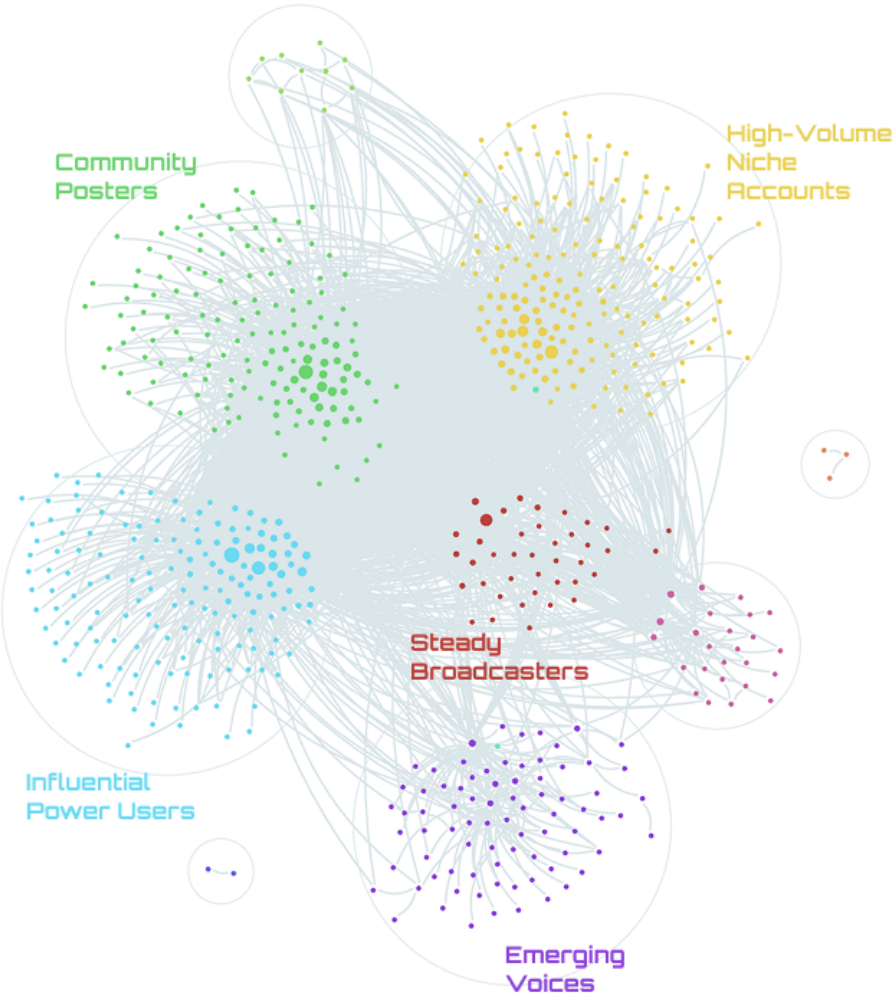


Figure 3 Five clusters of X accounts sharing TikTok video links within a 24-hour coordination window, as detected by CoorTweet (image export)

Using the CoorTweet tool, researchers detected both short-term (20-minute) and longer-term (24-hour) synchronisation, identifying five distinct campaigns, some resembling domestic influence operations, others showing signs of foreign interference (see Figure 3). While most accounts aligned with clear ideological camps, some shared contradictory material, mixing liberal and far-right narratives. Much of the content was politically charged, often critical of the government and Poland's geopolitical position.

These findings point to a fragmented yet coordinated ecosystem aimed at saturating the information space and shaping political debate ahead of the 2025 election, creating a hostile and uneven environment for actors who do not employ such tactics.

Overall, the activity across X and TikTok reflects a sustained influence operation, with infrastructure established well before the 2025 Polish presidential election and now actively exploited to distort the information environment. TikTok's share function played a central role in this dynamic: by enabling external reposting that feeds back into TikTok's own algorithmic recommendations, it amplifies visibility

while lowering barriers to manipulation. This design choice raises serious concerns about inadequate safeguards against misuse.

Table 13 presents the case study summary.

Table 13 Influence by Design: How Coordinated Networks Spread Political Content on Polish TikTok case study summary

Category	Element	***Detail options***
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☑ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☐ *Yes* ☐ *No* ☐ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: Brute force*
Data Collection	Method Used	☐ *API* ☑ *Scraping* ☑ *Third-party tool* ☑ *Manual*
	Platform / Tool	*e.g., Meta Content Library, CrowdTangle, Zeesschuimer, etc.*
	Data Collection Time Structure	☐ *Single snapshot* ☑ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☑ *Expert-driven* ☑ *Hashtag/topic-based* ☐ *Automated search*
Data Analysis	Type of Analysis	☐ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☑ *Mixed methods*
	Depth of Analysis	☐ *Surface-level* ☐ *Thematic coding* ☑ *Behavioural network analysis* ☐ *Other: _____*
Limitations & Open Questions	Identified Limitations	*Limited access to X and TikTok data forced reliance on scraping and unofficial APIs. The short study period prevented full analysis of behaviour of the cross-posting accounts, and with tools like Botometer now closed, confirming whether they are bots requires further investigation.*
	Key Open Questions	*cross-platform CIB methodology development*

5.2 AI-Generated Images of Pope Francis on Facebook Drive Traffic to Network of Suspicious Websites in Coordinated Campaign

5.2.1 Rationale

In vera.ai's ongoing efforts to monitor coordinated accounts discovered by the Vera AI Alerts[35], a team of researchers at the University of Urbino uncovered a significant disinformation campaign on Facebook pertaining to Pope Francis' health. This single operation generated thousands of posts a week, beginning on February 22, 2025. The campaign capitalised on the widespread attention surrounding the hospitalisation of the Pope, leveraging public interest and concern to disseminate misleading content. The investigation initiated by this case led to the identification of a wider network of accounts that regularly exploits large unmoderated Facebook groups to create very low quality content whose unique goal is to host a link, posted in the first comment of the post, to a set of potentially dangerous websites.

5.2.2 Methodology

For this case study we employed the Meta's Content Library (MCL) API via the Research Platform (the clean room environment directly maintained by Meta). The goal was to put the MCL to test on an iteration (starting from a seed of accounts to discover a wider network) as such iterations are the fundamental building block of the workflow designed to track coordinated activities over time developed in the context of vera.ai (Giglietto et al., 2023).

The data gathering pipeline started with a search query for all the posts in the library matching the query "Breaking& News & His & Holiness & Pope & Francis & has & failed" and was created between February 22 and March 10 2025. This query returned 133,777 posts. We used CooRTweet with restrictive parameters (time internal 10 seconds and edge weight 0.75) on this dataset of posts to perform a coordinated sharing behaviour detection analysis. The analysis returned a list of 701 coordinated surfaces (Pages or public groups). For each of these surfaces we collected all content posted during the last 6 hours (the original timeframe used for an iteration in the workflow). The resulting dataset consisted of 1,389 posts. Given the observed behaviour on Pope Francis's posts of posting the link to an external website in the first comment of the post we retrieved all the 617 first comments from these posts. We then extracted the link from the comments and returned a comprehensive list of domains employed in this information operation.

[35] Cases described in paragraphs 5.2, 5.3, 5.4, and 5.5 were discovered by the Vera AI Alerts. The Vera AI Alerts implements a 9-step workflow introduced in Giglietto et al. (2023). A detailed description of the Vera AI Alerts implementation is available at D4.2 Coordinated sharing behaviour detection and disinformation campaign modelling methods.

5.2.3 Main findings

The campaign's posts, while numerous, were strikingly uniform in their content. Each post contained identical text, accompanied by a visual collage (see Figure 4) comprising both AI-generated and authentic images of the Pope. These collages were likely created to circumvent automatic detection systems.

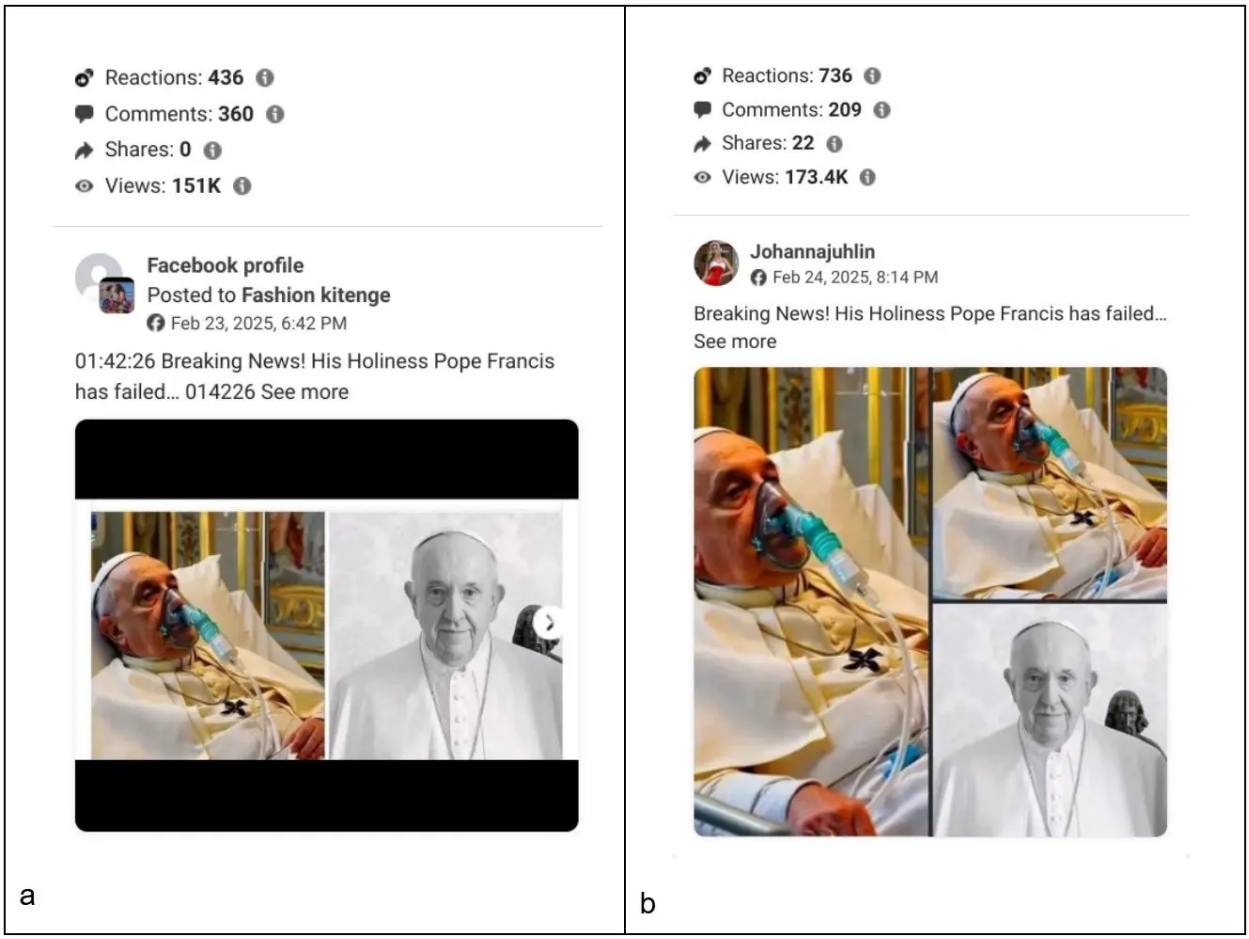


Figure 4 Visual collage comprising both AI-generated and authentic images of the Pope

Although one specific image within these collages was confirmed as AI-generated by the elDetector project of Univision Noticias[36], their fact-check label[37] has only been applied to a small fraction of the total posts.

Additionally, thanks to support from Agence France-Presse fact-checking unit, we located an early version of one of the AI images[38] featured in these collages that still display the Grok xAi logo (see Figure 5). The

[36] https://www.univision.com/noticias/foto-papa-francisco-mascarilla-oxigeno-creada-con-ia
[37] https://www.facebook.com/MasahuatTv/posts/papa-francisco-sigue-muy-delicado-de-saludel-vaticano-expresa-que-se-encuentra-m/1183563893777469/
[38] https://archive.ph/0SeqN

campaign's strategy involved posting a link in the first comment of each content, typically by the same user who created the original post.

Figure 5 AI-generated image of the Pope with logo (bottom right corner)

The links below direct users to a network of similar-looking breaking news blogs, all sharing the same template and tracking scripts (Table 14).

Table 14 A sample of website domains in the network (link to urlscan.io analysis)

sumary[.]store	premuimnew[.]store	mnews999[.]com
freshusanews[.]store	lunamedi3[.]info	toptreading[.]xyz
kingideas[.]space	taron[.]store	ponha[.]store
bknews168[.]store	worlds-recipes[.]online	diigital[.]press

Operations like these demonstrate how easily Meta's Facebook policies against false news, synthetic images, and links to potentially harmful websites can be circumvented. This case also reveals the limitations of Meta's AI systems in effectively identifying near-duplicate versions of previously fact-checked images. These findings are particularly concerning given the phasing out of the third-party fact-

checking programme in the United States, leading to non-detection of harmful posts which can easily exploit vulnerable users of the platforms.

Table 15 presents the case study summary.

Table 15 AI-Generated Images of Pope Francis on Facebook Drive Traffic to Network of Suspicious Websites in Coordinated Campaign case study summary

Category	Element	***Detail options***
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☐ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☑ *Yes* ☐ *No* ☐ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: _____*
Data Collection	Method Used	☑ *API* ☐ *Scraping* ☐ *Third-party tool* ☐ *Manual*
	Platform / Tool	*Meta Content Library*
	Data Collection Time Structure	☑ *Single snapshot* ☐ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☐ *Expert-driven* ☐ *Hashtag/topic-based* ☑ *Automated search*
Data Analysis	Type of Analysis	☑ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☐ *Mixed methods*
	Depth of Analysis	☑ *Surface-level* ☐ *Thematic coding* ☐ *Behavioural network analysis* ☐ *Other: _____*
Limitations & Open Questions	Identified Limitations	
	Key Open Questions	*how Meta policies against false news, synthetic images, and links to potentially harmful websites can be circumvented*

5.3 Coordinated Gambling Promotion on Facebook

In our investigation of coordinated online gambling promotion, we examined a network of 223 Facebook public groups that we discovered through our vera.ai project's monitoring system. This network represents one of the largest documented cases of systematic visual manipulation on social media platforms that we have encountered in our research.

The scale of this operation was unprecedented in our experience: these groups collectively published over 10,000 posts per hour on average, with membership sizes ranging from approximately 32,000 to over 612,000 members per group. The network's reach extended across multiple linguistic and cultural communities, with targeted content in Filipino, Urdu, and English, demonstrating the sophisticated audience segmentation strategies we observed.

Our case study centred on analysing over 2,300 distinct images embedded in posts from these groups between January 2017 and September 2024. What made this case particularly significant for our research was the dramatic transformation we observed following the public launch of ChatGPT in November 2022, which coincided with an exponential increase in both posting volume and visual sophistication. Monthly post volumes surged from an average of 2,121 to peaks exceeding 10 million posts, a 13,242% increase that fundamentally altered what we were seeing in the information ecosystem surrounding online gambling promotion.

The network's activities exemplified how coordinated actors exploit regulatory asymmetries in platform governance that we've been tracking. While Meta's policies strictly regulate paid gambling advertisements – requiring prior authorisation, age restrictions, and jurisdiction-specific compliance – organic content related to gambling faces significantly more permissive oversight. This regulatory gap creates what we identified as a structural loophole whereby gambling promotional material can circulate widely without triggering the oversight mechanisms applied to formal advertising, essentially functioning as advertising while appearing to be authentic user-generated content.

The gambling groups we studied represented a sophisticated ecosystem that targeted vulnerable populations through carefully crafted visual narratives that normalised gambling behaviours while exploiting psychological vulnerabilities and cultural sensitivities. This network provided us with a unique opportunity to examine how traditional marketing strategies evolved into AI-amplified manipulation architectures that operate at industrial scale.

5.3.1 Rationale

We selected this gambling network case study for multiple compelling reasons that address critical gaps we identified in understanding digital manipulation and platform governance.

Platform Architecture Exploitation: Through our analysis, we revealed how sophisticated actors understand and exploit the regulatory asymmetries built into social media platforms. While gambling advertisements face strict oversight requiring explicit authorisation and compliance with age restrictions, organic posts promoting gambling encounter minimal barriers. This asymmetry enables harmful promotional content to proliferate under the guise of peer-driven engagement, representing what we see as a fundamental challenge to current platform governance models that rely on content-type distinctions rather than impact-based assessments.

Public Health Urgency: Our research focus on online gambling promotion stems from its disproportionate impact on vulnerable populations, particularly young users and individuals with gambling problems, who are especially susceptible to algorithmically amplified persuasive content. The gambling industry's documented public health risks, combined with its rapid digital expansion and sophisticated targeting capabilities that we observed, made this a critical area for our intervention research. The normalisation

of gambling through social media platforms represents a significant threat to public health that existing regulatory frameworks struggle to address.

AI Technology Weaponisation: The emergence of generative AI tools offered us a unique natural experiment to examine how synthetic media technologies are being systematically weaponised to enhance manipulative content. The temporal alignment between ChatGPT's public launch and the exponential growth in posting volume that we documented provided compelling evidence of how rapidly emerging technologies can be co-opted for harmful purposes. This case allowed us to study the dual-use nature of AI technologies that can serve both creative and manipulative functions.

Methodological Innovation Opportunity: This case presented us with an ideal testing ground for developing new analytical approaches that combine large language models with traditional qualitative content analysis. We recognised that existing approaches to visual content analysis were inadequate for handling both the scale (thousands of images) and the interpretive complexity (cultural and connotative meanings) required for our investigation. The case offered an opportunity to pioneer scalable methods for analysing visual persuasion at scale.

Cultural and Linguistic Targeting: The network's sophisticated use of cultural localisation that we uncovered – including targeted content in Filipino leveraging peer group dynamics and Urdu content exploiting traditional family narratives – revealed how gambling promotion intersects with broader patterns of cultural manipulation and identity-based targeting that we've been studying. This provided insights into how harmful content adapts to exploit specific cultural vulnerabilities while maintaining underlying persuasive architectures.

Coordinated Behaviour Documentation: The case offered us an opportunity to document and analyse coordinated inauthentic behaviour in real-time, contributing to broader understanding of how influence operations function across social media platforms. The systematic nature of the coordination we observed, combined with the volume of content produced, provided rich data for understanding how manipulation scales in digital environments and how coordinated networks evolve their strategies over time.

5.3.2 Methodology

A mixed-methods approach was employed that combined computational techniques with qualitative analysis to handle both the scale and interpretive complexity of visual persuasion analysis.

Coordinated Behaviour Detection: the gambling network was identified through the Vera AI alerts workflow briefly described in section 5.2.2. The gambling network examined was the largest and most active, composed of 223 public groups specifically dedicated to gambling promotion.

Content Extraction Pipeline

The research leveraged Meta's Content Library (MCL), a transparency tool providing access to real-time public content datasets from Facebook, Instagram, and Threads. Using the MCL's producer list functionality, a comprehensive list of the 223 identified gambling groups was created and all posts published between January 13, 2017, and September 7, 2024 were extracted.

While the network produced over 70 million posts during this period, only 10,671 met MCL's criteria for downloadable content (requiring groups with 15,000+ followers). To extract multimedia content, a custom browser-based script was developed that automated the download of 2,323 distinct images while maintaining compliance with MCL's research data agreement requirements.

AI-Assisted Visual Analysis Pipeline

The methodology's innovation lay in using OpenAI's GPT-4o to generate both denotative (literal, objective) and connotative (cultural, emotional) descriptions of all images. This dual approach captured both surface-level visual elements and deeper cultural meanings, enabling more nuanced analysis of persuasive strategies.

Specific prompts were designed including: "Describe the connotative meaning of this image. Avoid phrases like 'the image conveys' or 'the image portrays'. Start the description without any introductory phrases. YOU MUST enclose the description within <desc> and </desc> tags."

The textual descriptions were transformed into vector embeddings using OpenAI's text-embedding-3-small model, then processed through UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction and HDBSCAN (Hierarchical Density-Based Spatial Clustering) for density-based clustering. This process generated 51 connotative clusters and 101 denotative clusters, with noise clusters containing 390 and 258 images respectively.

To identify persuasive drivers, co-occurrence matrices of images appearing in both connotative and denotative clusters were constructed. The premise was that specific combinations of connotative and denotative descriptions would indicate images depicting particular persuasion strategies. The 366 cluster combinations were ranked by frequency, with manual analysis focusing on combinations containing more than six images.

Qualitative Content Analysis

The team of 4 researchers involved in conducting the study participated in systematic qualitative analysis, examining each significant cluster combination until thematic saturation was reached. Each researcher independently analysed 21 cluster combinations, documenting emergent patterns and formulating descriptive labels for underlying gambling drivers.

Through consensus-building discussions, individual analyses were consolidated and interpretive discrepancies were resolved. This collaborative approach enhanced analytical rigor while accommodating the nuanced interpretation required for cultural and visual elements.

To examine generative AI's influence, multiple analytical approaches were employed. Quantitative analysis used Meta Content Library API data to track monthly posting volumes, with ChatGPT's November 30, 2022 launch as the intervention point. Statistical significance was assessed through t-tests, Wilcoxon rank-sum tests, regression models with interaction terms, and structural break detection.

Qualitative analysis examined visual content before and after ChatGPT's introduction, focusing on how established persuasion drivers were being executed in probable AI-generated content. Elements including visual style, composition, colour schemes, character representation, and narrative complexity were analysed to determine how generative technologies intensified existing persuasion tactics.

5.3.3 Main Findings

The analysis revealed seven core persuasive drivers included in a multi-layered system of visual strategies operating across the gambling network. These persuasive drivers are listed and described below.

1. Aspirational Wealth & Hyper-Masculine Status Fantasies (Present in 55% of combinations analysed)

This dominant strategy deployed luxury signalling through Mercedes and BMW vehicles, Rolex watches, and cash displays to link gambling participation with wealth accumulation and masculine power. The visual narratives constructed head-to-head competitive scenarios mimicking viral social media challenges, positioning gambling as a pathway to social recognition and dominance.

Particularly notable was the strategic use of sexualised imagery, featuring Asian models in revealing clothing positioned adjacent to jackpot announcements and bonus offers. These visuals were carefully situated in aspirational environments – lavish settings with luxury items – suggesting gambling as access to a sophisticated lifestyle. The imagery reflected a broader masculine imaginary where power, control, and success were associated with the capacity to attract and display beauty and wealth.

2. Manufactured Trust Through Transactional Proofs (37% of combinations analysed)

The network systematically employed financial documentation to establish credibility and mitigate user scepticism. Screenshots of successful withdrawals, complete with timestamps and reference numbers, were paired with celebratory messages like "Transfer successful" and "Withdrawal completed." The analysis showed the use of recognisable payment intermediaries, particularly GCash (a popular Philippines payment app), fostered perceptions of legitimacy and transparency.

Crucially, the network employed humour as a psychological buffer against gambling losses. Memes and humorous content referenced gambling failures, transforming individual financial loss into shared cultural experiences that legitimised continued engagement. This strategy reframed losses as normal occurrences to be met with humour rather than reflection, encouraging habitual betting behaviour.

3. FOMO & Urgency Conditioning

Dynamic countdown timers, flashing "claim" buttons, and time-limited offers created artificial scarcity contexts designed to exploit well-documented cognitive biases. The interface design deliberately emulated slot machine environments using bright gradients, animated icons, and reward escalation mechanisms to produce continuous reward anticipation cycles.

This architecture reflected principles of operant conditioning, where progress indicators, incremental bonuses, and multi-level reward ladders acted as reinforcing stimuli encouraging repetitive gambling behaviours. The systematic alternation between rewards and small reinforcements embedded gambling into habitual digital routines, making disengagement increasingly difficult as users invested more time and resources.

4. Cultural & Linguistic Localisation

The analysis demonstrated sophisticated understanding of cultural targeting, with 56 clusters showing Filipino and Urdu language content. Filipino-language materials integrated gambling promotion with

culturally specific symbols, local slang, and peer pressure dynamics, facilitating naturalisation of gambling as community practice through familiar cultural motifs.

The Urdu-language cluster revealed particularly complex ideological messaging, dominated by emotionally charged depictions of women in distress, domestic conflicts, and traditional role performance while male figures appeared distant or abusive. This content exploited collective emotional and moral sensibilities, reinforcing cultural conservatism while subtly embedding gambling platforms within broader moral narratives about family values and social cohesion.

5. Gamification & Low Entry Barriers

Extensive gamification techniques were documented that replicated slot machine logic in mobile environments. Spin buttons, multipliers (10x, 20x), visual representations of quests (temples, treasures, adventures), and escalating reward systems created immersive, overstimulating experiences designed to make gambling appear approachable and risk-free.

Low-entry incentives like "$2 free play" or "₱0.50 bets" were prominently displayed alongside screenshots of large winnings, creating illusions that small investments could yield disproportionately large returns. This exploited the gambler's fallacy and illusion of control, fostering continued engagement despite repeated losses.

6. Celebrity & Influencer Endorsements

The analysis showed public figures lent credibility and aspirational status to gambling brands through professionally designed promotional layouts. National sports figures, particularly prominent athletes, leveraged affective bonds of national identity and masculine ideals.

The endorsements transferred symbolic capital from public figures to gambling brands, positioning participation not merely as acceptable but as aspirational and patriotic. Graphics blended corporate polish with emotional appeal through bold fonts, high-contrast colours, and celebratory poses that conveyed excitement while offering reassurance through familiar, trusted endorser images.

7. Social Relations Exploitation

The network systematically transformed users into active recruiters and content amplifiers. "Invite friends" mechanisms and reward-based sharing incentives encouraged viral dissemination while turning players into promotional agents. The convergence of aspirational imagery, emotional manipulation, urgency mechanisms, and viral strategies created self-reinforcing ecosystems of attention capture and behavioural conditioning.

Transformation of Posts Content and Operation Scale with Generative AI

Unprecedented changes were documented following ChatGPT's November 2022 launch, providing compelling evidence of how generative AI technologies transformed gambling promotion operations.

Statistical analysis revealed consistent shifts in posting behaviour. Mean monthly posts increased from 2,121 to 280,952 (13,242% increase), while median posts rose from 30 to 800,000 (2,666,566% increase). Regression analysis confirmed this effect was statistically significant ($p<0.0001$), demonstrating both immediate level increases and significant acceleration in growth trajectory.

The pre-ChatGPT posting trend was essentially flat (154.9 posts per time unit, p=0.973), while post-ChatGPT slope increased dramatically (481,155 posts per time unit). Structural break analysis detected a significant breakpoint in July 2023, suggesting delayed but profound impact as generative AI technologies were increasingly integrated into promotional strategies.

AI-generated content displayed consistent aesthetic markers distinguishing it from traditional promotional materials. These included hyper-realistic lighting gradients, unnaturally smooth surfaces, and near-perfect symmetry rarely achieved through manual production (see Figure 6left). The aesthetic combined saturated, dreamlike colour schemes with densely layered compositions featuring improbable arrangements of sharks, treasure chests, human figures, and jackpot cues.

The most viewed AI-generated posts exemplified this transformation. One prominent example featured a massive shark emerging from underwater depths, surrounded by poker chips and additional sharks, with a slot machine positioned centrally. The image advertised "$20 FREE PLAY" with "NO TASKS" in bold, glowing typography, merging slot machine mechanics with fantasy gaming and high-risk adventure narratives.

Figure 6 On the left, side-by-side display of luxury cars used to associate gambling with wealth, competition, and masculine status. On the right, a woman in lingerie positioned next to promotional gambling content, exemplifying the use of sexualised femininity to enhance aspirational appeal.

Another example featured polished, idealised women at casino tables with confident gazes, surrounded by soft-focus lighting and luxurious interiors (see Figure 6right). The image promoted "REDEEMABLE FREEPLAY" and "maximum cashout" with urgent messaging. These posts achieved massive engagement: one variant garnered 4.3 million views, 160 reactions, 1,100 comments, and 1,000 shares, while another received 3.3 million views, 596 reactions, 3,300 comments, and 1,500 shares.

AI-generated content didn't replace existing persuasion drivers but enhanced and intensified them through improved visual immersion and narrative complexity. The technology enabled construction of imaginative universes where reward systems were not only emotionally appealing but symbolically elevated, creating more cognitively captivating experiences while maintaining underlying psychological manipulation techniques.

Table 16 presents the case study summary.

Table 16 Coordinated Gambling Promotion on Facebook case study summary

Category	Element	***Detail options***
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☑ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☐ *Yes* ☐ *No* ☑ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: _____*
Data Collection	Method Used	☑ *API* ☐ *Scraping* ☐ *Third-party tool* ☐ *Manual*
	Platform / Tool	*Meta Content Library, CrowdTangle*
	Data Collection Time Structure	☐ *Single snapshot* ☑ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☐ *Expert-driven* ☑ *Hashtag/topic-based* ☐ *Automated search*
Data Analysis	Type of Analysis	☐ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☑ *Mixed methods*
	Depth of Analysis	☐ *Surface-level* ☑ *Thematic coding* ☑ *Behavioural network analysis* ☐ *Other: _____*
Limitations & Open Questions	Identified Limitations	*Platform API restrictions (MCL limited to popular accounts), reproducibility challenges due to CrowdTangle deprecation, incomplete network coverage*
	Key Open Questions	*Long-term behavioural effects on users, complete amplification mechanisms, comprehensive detection methods for AI-generated content*

5.4 Coordinated Visual Propaganda in Pro-Putin Facebook Groups

This study examines a coordinated network of pro-Putin Facebook fan groups identified through the vera.ai workflow, which detected coordinated networks via CrowdTangle. The initial detection identified 27 groups, with 15 included in the final study (averaging 45,586 followers per group). Notable groups included "Putin's Team" (95,890 followers), "Vladimir Putin is the best president" (83,379 followers), and "WorldRusWorld" (87,380 followers).

The 15 groups had approximately 55,000 publicly available posts in total. Over the one-year observation period (24 April 2024–23 April 2025), researchers downloaded 5,917 posts, maintaining sustained coordinated activity analysis. Each post garnered an average of 27 user interactions (likes, comments, shares), indicating consistent audience engagement. The researchers extracted 1,089 unique visuals (static images and keyframe videos) from these posts for detailed analysis.

The study focused on understanding how visual propaganda strategies were employed across this coordinated network to shape political discourse and build ideological alignment. The timing proved particularly significant, with a notable spike in posting activity during November-December 2024, coinciding with Donald Trump's re-election as U.S. President.

5.4.1 Rationale

The study addresses several **critical gaps in digital propaganda research.** First, Facebook remains significantly understudied compared to Twitter, despite having a larger and more diverse user base. Most academic research has disproportionately focused on Twitter due to its more accessible API, leaving Facebook's role in hosting coordinated influence operations relatively unexplored.

Second, there is limited research on visual framing analysis specifically examining content shared by coordinated Facebook networks. This gap is particularly concerning given the increasing recognition of visual content as a primary tool of influence on platforms where images and videos are algorithmically prioritised in user feeds.

Third, multilingual propaganda efforts across national and cultural boundaries remain underexplored, even though they play a strategic role in Russia's geopolitical influence campaigns. The visual nature of propaganda allows it to bypass language barriers and content moderation systems more effectively than text-based content.

The study is particularly timely given the evolution of Russian influence operations since the 2014 annexation of Crimea. These operations have grown in sophistication, combining overt state media channels with covert coordinated inauthentic behaviour (CIB) tactics. The researchers positioned their work within the broader context of Foreign Information Manipulation and Interference (FIMI) - hostile actions that intentionally distort information environments to achieve geopolitical goals.

The visual focus was justified by research showing that visual content not only bypasses language barriers but also heightens emotional salience, making it a powerful vehicle for disinformation and ideological messaging. As generative AI becomes integrated into visual production, understanding these propaganda mechanisms becomes increasingly urgent for platform accountability and regulatory responses.

5.4.2 Methodology

The study employed a mixed-methods approach combining automated detection systems with innovative human-AI collaborative analysis. The network was initially identified using the vera.ai workflow, a European research tool designed to detect coordinated inauthentic behaviour on Facebook using CrowdTangle data. The vera.ai workflow detected coordinated networks via CrowdTangle, with initial detection identifying 27 pro-Putin groups, of which 15 were included in the final study. Groups had approximately 55,000 publicly available posts in total, with 5,917 posts downloaded for analysis during the specified timeframe. Account lists were imported into Meta Content Library for structured data retrieval, and 1,089 unique visuals were extracted using the MCL Media Downloader tool. For videos, R scripts extracted first keyframes for analysis.

The researchers adapted a visual framing model from Rodriguez and Dimitrova (2011) and Chakraborty and Mattoni (2023), focusing on four key dimensions. The denotative dimension examined main visual elements including human elements, non-human elements, and text. The stylistic dimension analysed visual composition through foreground, background, and colour palette. The connotative dimension explored underlying meanings through figurative symbols, emotions invoked, and underlying ideas. The contextual dimension considered the broader public debate context.

The study introduced an innovative methodology using ChatGPT-4o for systematic visual analysis. The first 10 images were manually coded to refine the analytical framework, followed by 50 images comprising 25 static and 25 video keyframes analysed using AI assistance. Each image was individually uploaded and coded using the structured dimensions, with researchers providing iterative feedback to ensure accuracy and nuance. The AI demonstrated particular strength in multilingual text translation and descriptive detail. The analysis is ongoing at the time of preparing this report.

However, limitations of AI analysis included occasional failure to recognise political figures less frequently represented in Western media, such as Russian Foreign Minister Sergey Lavrov and Iran's President Masoud Pezeshkian. This highlighted the model's reliance on English-language information sources.

5.4.3 Main Findings

The analysis revealed three primary visual frames employed across the coordinated network. The first was **glorified leadership**, where Vladimir Putin consistently appeared as a mythologised figure portrayed as a heroic statesman, military commander, paternal figure, and mythic warrior. Visual signifiers included Russian flags, military uniforms, religious symbolism, and pristine natural settings. These portrayals reinforced alignment between state power and personal charisma.

The second frame **emphasised military strength and technological prowess**, with Russian armed forces and weaponry showcased with precision and grandeur, often through immersive first-person perspectives or gaming-influenced cinematic aesthetics. This framing emphasised operational competence and intimidation through spectacle. The third frame invoked **historical and civilisational exceptionalism** through Soviet-era nostalgia, space imagery, WWII symbols like the Saint George's Ribbon, and references to Russia's aerospace legacy. Arctic frontiers, military parades, and cultural motifs asserted timeless Russian greatness and moral superiority.

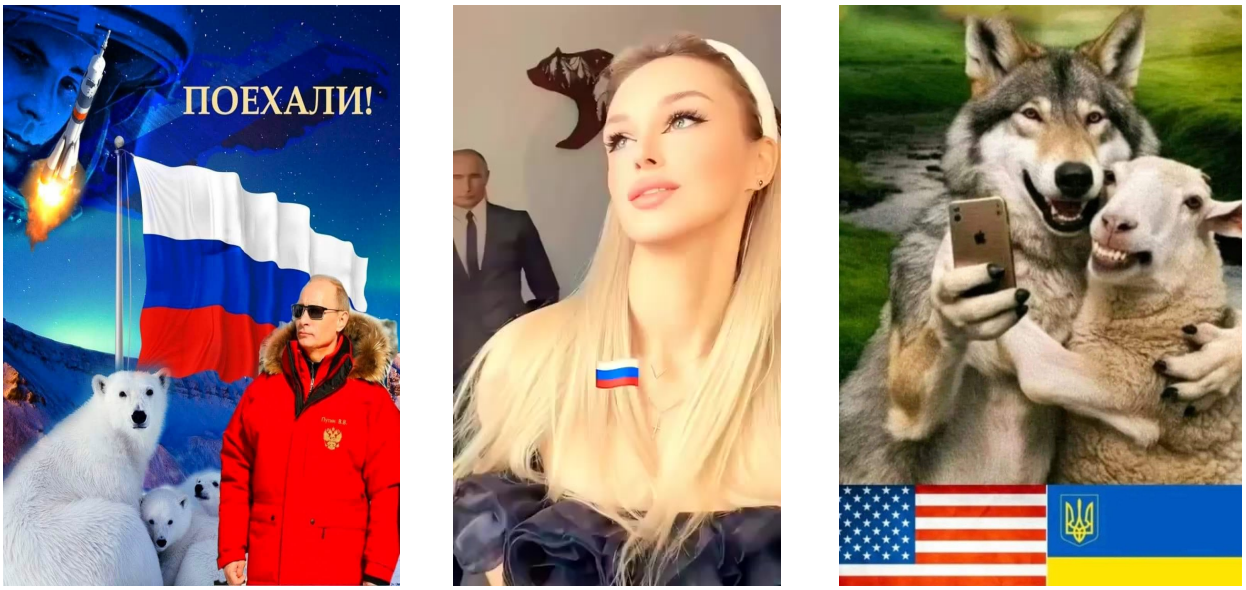
Antagonistic framing consistently depicted Western leaders, particularly Volodymyr Zelensky and Joe Biden, through mockery, satirical imagery, and infantilising portrayals. Visual metaphors included predator-prey relationships with wolves versus sheep, using symbolic animals to represent geopolitical power dynamics.

The study revealed significant differences between static and dynamic visual formats. Static images focused on symbolic density and message clarity, using portraits, photomontages, and nationalist tableaux. They emphasised Putin as a central icon surrounded by national symbols, employed high compositional control for emotional resonance, and utilised humour and satire through meme formats. Dynamic visuals emphasised motion, immediacy, and immersion, featuring first-person military perspectives, rocket launches, and parade sequences. They simulated presence and experience rather than relying solely on symbols, optimised for visceral impact and performative legitimacy, and conveyed urgency, realism, and operational competence.

The research revealed a **comprehensive visual propaganda ecosystem designed to emotionally condition audiences through coordinated visual codes**. Static images served to assert identity and belonging, while dynamic visuals simulated power and control. Together, these formats created what Graham (2025) termed "infrastructure of truth" – a performative mechanism asserting authenticity through visual spectacle.

The study found clear alignment with known Russian strategic narratives but revealed a shift toward **greater visual personalisation and affective storytelling** (see Figure 7). Unlike purely deceptive operations, this campaign emphasised emotional branding and character-centric visuals, moving from confusion-based tactics to conversion through entertaining content.

The study demonstrated the **viability of human-AI collaboration in large-scale visual content analysis**, offering a scalable approach for future research. The prompt-assisted methodology proved particularly valuable for multilingual content analysis and systematic visual framing identification. The findings contribute to understanding how algorithmic curation, visual semiotics, and media infrastructure interact in contemporary influence operations, calling for urgent rethinking of platform accountability and regulatory approaches to visual-first propaganda in the digital age.



(a) (b) (c)

Figure 7 Images circulated in the fan groups that reveal the type of messaging common in these groups: a) Putin as a hero; b) Putin's popularity among the youth; c) Ukraine's dependency on the west (in the image, US is seen as the wolf posing with the sheep, Ukraine)

Table 17 presents the case study summary.

Table 17 Coordinated Visual Propaganda in Pro-Putin Facebook Groups case study summary

Category	Element	*Detail options*
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☐ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☑ *Yes* ☐ *No* ☐ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: _____*
Data Collection	Method Used	☑ *API* ☐ *Scraping* ☐ *Third-party tool* ☐ *Manual*
	Platform / Tool	*Meta Content Library*
	Data Collection Time Structure	☑ *Single snapshot* ☐ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☑ *Expert-driven* ☐ *Hashtag/topic-based* ☑ *Automated search*
Data Analysis	Type of Analysis	☐ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☑ *Mixed methods*
	Depth of Analysis	☐ *Surface-level* ☑ *Thematic coding* ☐ *Behavioural network analysis* ☐ *Other: _____*
Limitations & Open Questions	Identified Limitations	*scalability of the method in the context of platform changes/updates/data access limitation*
	Key Open Questions	*the viability of human-AI collaboration in large-scale visual content analysis*

5.5 A Longitudinal Analysis of Coordinated Pro-Bolsonaro Facebook Accounts

This study examines how major endogenous and exogenous events influence affective engagement patterns in a hyperpartisan pro-Bolsonaro network of Facebook accounts. Hyperpartisan online communities are often characterised as impermeable echo chambers with predictable engagement patterns, yet this assumption has rarely been tested longitudinally. Political participation through digital spaces is particularly relevant in Brazil, where the country experienced a significant rise in political polarisation with Bolsonarism that culminated in the 2023 attempted coup.

This is the first study to longitudinally analyses interaction patterns in Brazilian polarised communities over three years (2021-2023). We employed a novel mixed-methods approach combining semi-supervised computational content analysis through LLMs, time series analysis, and statistical analysis of interaction

patterns to examine over 12 million posts from 53 groups and 4 pages. We constructed two behavioural indices: the Emotional Polarisation Index, measuring love-versus-angry reactions, and the Engagement Balance Index, capturing comment-versus-share interactions.

The study found that hyperpartisan communities displayed surprisingly unstable engagement patterns over time, contradicting traditional theories that assume polarisation as consistent. A major shift occurred in 2023 around Lula's inauguration and the January coup attempt, marked by declining emotional intensity, increased internal debate, and weakening opposition responses. These findings demonstrate that user behaviour in hyperpartisan groups was driven by current events and context in addition to ideological positions. The study contributes methodologically through its longitudinal approach and use of machine learning for political actor recognition, offering new insights into the dynamic nature of political polarisation in digital environments.

5.5.1 Rationale

Political participation aimed at establishing extreme ideologies and potentially subverting democratic institutions has become increasingly prominent within hyperpartisan online communities (Marwick & Lewis, 2017). While considerable attention has focused on the United States, less investigated countries from the Global South deserve scholarly attention given their unique political contexts (Fenoll et al., 2024).

Brazil represents a compelling case as the world's fifth-largest digital market, where social media has reshaped political engagement (Ruediger & Grassi, 2022). Political participation increasingly occurs within hyperpartisan Facebook networks that exhibit ideological segregation and disinformation consumption patterns (Recuero et al., 2022). The Bolsonaro era paralleled the Trump presidency, with both following similar trajectories culminating in attempted coups upon leaving office (Bastos & Freitas, 2025). Both leaders relied on anti-institutional rhetoric, emotionally charged narratives, and direct social media engagement, fostering hyperpartisan communities that promoted institutionally distrustful engagement (Bastos & Recuero, 2023; Chaguri & do Amaral, 2023).

Within these digital environments, exposure to opposing viewpoints rarely results in deliberation but instead reinforces partisan divisions and facilitates misinformation spread, with ideological confrontation frequently marked by incivility (Guimarães et al., 2021; Rossini et al., 2021). Understanding these dynamics is crucial for assessing threats to democratic discourse and institutional stability.

Studies on Brazilian political polarisation in digital media have increased substantially, yet two relevant gaps persist. First, the longitudinal dimension: studies focus on single crisis events, missing the broader temporal context of social media participatory behaviours. Second, analytical scope: existing research typically observes attitudes toward in-group and out-group content in cross-cutting spaces like news comment sections, leaving unexplored how these dynamics manifest across different digital environments and interaction modalities. We employed the methodological pipeline described below to address these gaps.

5.5.2 Methodology

This study examines a pro-Bolsonaro community comprising 53 Facebook groups and 4 pages identified through Vera AI Alerts. Using CrowdTangle, we collected 12,153,915 posts from January 2021 to December 2023.

We constructed two behavioural indices as engagement proxies. **The Emotional Polarisation Index (EPI)** measures emotional response balance: (love reactions - angry reactions) / (love reactions + angry reactions), ranging from -1 (purely angry) to +1 (purely love). **The Engagement Balance Index (EBI)** captures amplification versus discussion: (shares - comments) / (shares + comments), ranging from -1 (comment-dominated) to +1 (share-dominated).

Both indices showed **high boundary clustering** (94.3% and 66% at extremes, respectively) and temporal volatility. Using a 30-day rolling window, we identified 96 volatile days (8.8% of observations) when the standard deviation exceeded the 95th percentile (see Figure 8). These clustered into seven contiguous instability periods lasting 3-33 days, coinciding with major political events like the January 2023 attempted coup.

Figure 8 Temporal dynamics of social media engagement measures for pro-Bolsonaro content. Daily means of EPI (top panel) and EBI (bottom panel) with high volatility periods highlighted (dots) and major political events marked (vertical lines)

We filtered 712,287 posts with textual content and developed a multi-label classification system using fine-tuned GPT-4o to identify six political actor categories: (1) Bolsonaro entities, (2) Lula entities, (3) Supreme Court/institutions, (4) Mainstream media, (5) Armed forces, and (6) Other political entities. The model achieved 67.71% exact match accuracy and 0.85 F1-score on validation data (n=449).

We estimated multinomial logistic regression models predicting EPI and EBI categories (using tertile breakpoints) based on political actor presence. Models included binary indicators for five actor categories and controlled for total engagement and emotional reaction volumes. Coefficients were interpreted as relative risk ratios indicating associations between political actors and user engagement patterns during periods of affective instability.

5.5.3 Main Findings

This analysis examined engagement patterns across seven critical political timeframes in pro-Bolsonaro Facebook groups, revealing unexpected deviations from traditional partisan dynamics.

Bolsonaro-Related Content

Contrary to expectations of stable in-group support, Bolsonaro-related content showed significant temporal variation. Love-dominated responses peaked during 2021 but declined dramatically by 2023, dropping from strong support during the attempted coup period to minimal positive engagement by December 2023. This decline suggests external events, COVID-19 mismanagement and the January 2023 insurrection, weakened traditional loyalty patterns rather than sustaining in-group solidarity.

Most notably, comment-dominated engagement revealed a fundamental shift, contradicting echo-chamber theories. All 2023 timeframes showed significant increases in argumentative responses, contrasting sharply with 2021's suppression of discussion. This transformation following Lula's electoral victory indicates evolving community dynamics where supportive amplification gave way to internal debate and disagreement.

Lula-Related Content

Lula references produced expected oppositional behaviour with notable temporal variations. Angry responses peaked early during Lula's political reengagement but dissipated entirely in later timeframes, contradicting assumptions about sustained oppositional mobilisation. Conversely, suppression of love reactions remained consistent but weakened progressively through 2023, possibly indicating intervention by anti-Bolsonaro users following the traumatic attempted coup event.

Share-dominated interactions followed the most predictable oppositional pattern, with consistent suppression across all timeframes that gradually weakened in 2023. This aligns with broader trends toward decreased oppositional intensity over time.

Institutional Actors

- **Armed Forces**: Content citing this actor generated the most volatile and inconsistent responses, with engagement patterns fluctuating unpredictably between suppression and promotion across periods. This instability reflects the military's unique symbolic position in Bolsonaro's political narrative, where community reactions are linked closely to specific real-world military actions rather than consistent partisan treatment.
- **Public Institutions**: Content citing this actor generally conformed to oppositional expectations, consistently suppressing supportive sharing behaviours. However, unexpected increases in comment-dominated engagement appeared during specific timeframes, suggesting temporary shifts toward argumentative engagement during governmental controversies rather than dismissive avoidance.
- **Mainstream Media**: Content citing this actor revealed the most complex and contradictory patterns. While early timeframes showed expected oppositional responses with anger promotion and love suppression, later periods demonstrated possible shifts toward more neutral emotional engagement. Most strikingly, certain periods showed increased share-dominated engagement, suggesting strategic amplification of media content that portrayed outlets negatively or supported pro-Bolsonaro narratives.

Conclusion

These findings fundamentally challenge traditional echo-chamber theories by revealing that partisan engagement operates through multiple dynamic mechanisms rather than stable patterns. Pro-Bolsonaro communities demonstrated contextually dependent responses that shifted significantly over time, particularly following traumatic political events like the attempted coup. Rather than maintaining consistent supportive or oppositional behaviours, these groups exhibited evolving engagement strategies influenced by major political developments, platform dynamics, and changing user participation patterns. The data suggests that partisan online communities are more fluid and responsive to external events than previously theorised.

Table 18 present the case study summary.

Table 18 A Longitudinal Analysis of Coordinated Pro-Bolsonaro Facebook Accounts case study summary

Category	Element	***Detail options***
Impact	Observed Impact	☑ *Yes* ☐ *No* ☐ *Inconclusive*
	Potential Impact	☐ *High* ☐ *Medium* ☐ *Low*
Amplification Evidence	Evidence of Algorithmic Amplification	☑ *Yes* ☐ *No* ☐ *Not Investigated*
	Type of Amplification (if any)	☐ *Recommendation system* ☐ *Trending* ☐ *Engagement* ☐ *Other: _____*
Data Collection	Method Used	☑ *API* ☐ *Scraping* ☐ *Third-party tool* ☐ *Manual*
	Platform / Tool	*Meta Content Library*
	Data Collection Time Structure	☐ *Single snapshot* ☑ *Longitudinal* ☐ *Continuous* ☐ *Experiment-based*
	Data Source Selection	☑ *Expert-driven* ☐ *Hashtag/topic-based* ☑ *Automated search*
Data Analysis	Type of Analysis	☐ *Automated (keyword/topic clustering)* ☐ *Manual coding* ☑ *Mixed methods*
	Depth of Analysis	☐ *Surface-level* ☑ *Thematic coding* ☐ *Behavioural network analysis* ☑ *Other: Statistical Inference*
Limitations & Open Questions	Identified Limitations	*platform restrictions, reproducibility, incomplete data, time constraints*
	Key Open Questions	*rethinking echo-chamber information propagation patterns*

6 Conclusions and Recommendations

6.1 Disinformation Impact Conclusions and Recommendations

The vera.ai tools can support impact measurement, particularly concerning content reach and amplification. These tools are best suited for quantitative assessments, where metrics such as cross-platform distribution, audience exposure, and engagement levels can be tracked. However, restrictions on APIs imposed by some platforms in recent years jeopardise the tools' ability to collect current data and stay up-to-date. Vera AI Alerts – which flag Facebook content whose amplification is being coordinated – has used the Meta Content Library since late 2024 (following the shutdown of CrowdTangle and with the limitations detailed in the deliverable D4.2), and some progress has been made in implementing the TikTok Research API. Nonetheless, X remains one of the platforms to which we currently lack API access (our application was denied).

However, the applicability of current tools is more limited when assessing qualitative dimensions of impact, such as psychological influence, behavioural change, or societal harm, largely due to the absence of measurable indicators. Although tools for detecting persuasion techniques[39] are under development, significant gaps remain. Future developments in vera.ai, alongside clear guidance on defining and measuring impact indicators, could help bridge these gaps, particularly in areas where existing frameworks fall short. vera.ai's potential to enhance qualitative assessment lies in its capacity to provide contextual analysis[40] and integrate credibility signals[41]. These features would enable more robust evaluations of behavioural changes prompted by disinformation or the interventions designed to counter it, thereby strengthening vera.ai's role in comprehensive impact assessment. Additionally, content analysis presents an often underutilised opportunity. Exploring elements like language, tone, format, and narrative structure can uncover how content shapes impact, highlighting who the content targets, the context in which it spreads, and the risks it may pose. This deeper analysis allows researchers to better understand the mechanisms behind amplification, emotional manipulation, audience segmentation, and the broader effectiveness of disinformation campaigns.

An effective impact assessment framework should account for:

- The negative effects of a disinformation or influence campaign, such as its reach, amplification, and societal harm.
- The positive effects of countermeasures, evidenced by a measurable decline in impact metrics like reduced visibility, engagement, or behavioural influence.

vera.ai tools can support both objectives, offering capabilities to track the spread and intensity of disinformation while also evaluating the effectiveness of responses aimed at mitigating its impact.

Additionally, the methodologies for disinformation impact assessment enable a more systematic investigation of disinformation dynamics, by mapping problematic source clusters and their influence on

[39] https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier
[40] https://cloud.gate.ac.uk/shopfront/displayItem/news-framing-classifier
[41] https://gatenlp.github.io/we-verify-app-assistant/supported-tools#credibility-signals

mainstream discourse, and by classifying coordinated behaviours across diverse actors and degrees of authenticity:

- **The Post-truth Space Mapping Technique Protocol** offers a way to detect and visualise clusters of problematic sources, highlighting how fringe communities and low-visibility actors influence mainstream discourse; it has proven valuable for fact-checkers in identifying high-impact disinformation sources.
- **The CIB Typology Development Technique** provides a mixed-methods framework to categorise coordinated behaviours across a spectrum of actors, revealing overlapping coordination patterns and a continuum of inauthenticity that challenges clear distinctions between legitimate and deceptive activity.

6.2 Addressing Amplification as a Design Risk: Conclusions and Policy Recommendations

Personalised recommendation systems on the platform's "For You" Page play a defining role in shaping online visibility, influencing user behaviour, and potentially exposing societies to systemic risks.

While the distinction between algorithms and algorithmic amplification is often blurred in public discourse, this analysis has shown that amplification constitutes not merely a technical output, but a socio-technical process with significant cultural and political implications.

The case studies reviewed confirm that amplification on TikTok functions primarily through implicit signals – such as watch duration – rather than explicit feedback, and that this leads to rapid personalisation. Over time, users are funnelled into highly interest-aligned, often narrow content loops. While this enhances user retention, it also reduces exposure to diverse perspectives and increases the risk of reinforcing echo chambers. These findings are consistent with growing evidence that algorithmic amplification can escalate polarisation and misinformation, particularly in sensitive democratic contexts.

Alarmingly, this dynamic has manifested in real-world disinformation campaigns during recent European electoral cycles, as well as the exploitation of amplification mechanisms for malicious purposes. The European Commission[42] opened a formal investigation into TikTok under the DSA after suspected interference during the **2024 Romanian presidential election**, which was ultimately annulled. Fact-checkers at Correctiv[43] identified dozens of viral videos on TikTok containing disinformation about the **2025 German federal election**. Despite being reported, many of these remained online – underscoring the platform's inadequate response mechanisms and opacity around content moderation and amplification criteria. During the **2025 Polish presidential election**, TikTok played a central role in the spread of false and misleading political narratives, including doctored videos and targeted disinformation aimed at discrediting opposition parties. An investigation by AI Forensics uncovered systemic amplification of such content, facilitated by TikTok's opaque recommendation logic.

[42] https://www.reuters.com/business/eu-opens-investigation-into-tiktok-over-election-interference-2024-12-17/

[43] https://correctiv.org/faktencheck/hintergrund/2025/02/21/melden-zwecklos-tiktok-ignoriert-desinformation-zur-bundestagswahl/

These cases underline that TikTok's recommendation system is not merely reflective of user behaviour, but actively constructs and reinforces narratives that can undermine democratic integrity. They also expose a **profound mismatch between the platform's public commitments under the DSA and its real-world performance in mitigating systemic risks**.

Despite acknowledging algorithmic systems as potential vectors of harm, TikTok's 2023 and 2024 DSA-mandated risk assessments fail to offer meaningful transparency. Their treatment of algorithmic amplification is limited to general content ineligibility rules, with no technical explanations, quantification of risks, or discussion of design-level mitigation. **The focus is mainly on content risks, while the role of recommendation systems is largely overlooked. This design-blindness allows harmful amplification to persist under the guise of user preference and free expression.**

To address these critical shortcomings, the following measures are recommended:

- **Ensure effective researcher access to data under Art. 40**. Vetted researchers should have meaningful access to TikTok's recommender systems and internal datasets. This includes safe, privacy-compliant mechanisms for auditing algorithmic decisions, tracking amplification trajectories, and understanding the impact of feedback loops on users over time.
- **Strengthen the substance of DSA transparency reports under Art. 42**. Platforms' risk assessment reports under the DSA must go beyond declarative statements. They should be made available in a machine-readable format and include empirical evidence, algorithmic audit trails, reproducible methods, and disaggregated data about content curation and reach. Clear distinctions must be drawn between what content is amplified and how it is amplified.
- **Acknowledge and address amplification as a systemic risk under Arts. 34 and 35**. While platform design is already recognised as a potential systemic risk, current practice tends to underestimate the role of algorithmic amplification. Greater emphasis should be placed on how engagement-driven recommendation systems can amplify harmful content, shifting the focus from content alone to the underlying architecture that determines its reach. When conducting systemic risk assessments, platforms should explicitly evaluate how their recommendation systems may be exploited by malicious actors – including through CIB, automated manipulation, or strategic content flooding.
- **Mandate algorithmic transparency and accountability, also in line with Art. 38**. Platforms should be required to disclose the key parameters of their recommendation engines, the variables that influence amplification, and the governance mechanisms for moderation and user control. Additionally, Art. 38 of the DSA mandates the option for non-profiling-based recommendation systems, reinforcing the need for genuine user control over their experience on the platform.

6.3 General Conclusions from Case Studies

6.3.1 TikTok Algorithmic Amplification

The two TikTok algorithmic-amplification case studies yield lessons about both the platform's recommendation logic and the practical limits of researching it. The first case study shows that **audit findings on the "For You" feed are highly unstable**. Platform updates, changing HTML structures, rising

bot-detection defences, regional variation and short-lived content shifts all undermine attempts to reproduce earlier results. Even over the few weeks required to run their agent-based sock-puppet experiment, the authors had to repeatedly patch code to keep pace with new ads, livestream formats and shadow bans. In parallel, weak documentation in prior audits – missing code repositories, undefined watch-time parameters and inconsistent study designs – further eroded comparability. The conclusion is straightforward: without longitudinal, fully documented, multi-platform audits that simulate real users on mobile as well as web, regulators and researchers cannot rely on point-in-time tests to track systemic risks or evaluate policy interventions. Moreover, when reproducible methods are applied, the strongest personalisation driver remains implicit engagement. Watch duration, especially full or repeated views, consistently narrowed content to niche, interest-aligned videos, whereas explicit feedback such as likes and follows left the recommendation mix virtually unchanged. Location mattered only briefly at session start-up. These observations reinforce the concern that TikTok's recommender system rewards passive attention signals, accelerating users into self-reinforcing loops while offering little agency through overt interaction.

The second case study provides an alternative facet of TikTok research. The comment-section study of Romania's 2025 presidential run-off extends these insights from feed curation to downstream **user behaviour**. The investigation revealed signals of inauthentic coordination, including widespread repetition of identical or near-identical comments and synchronised posting patterns. A small group of fewer than 3,000 "power commenters" produced nearly half of all comments, often concentrating activity on specific candidates in a way that suggested targeted campaigning. These dynamics indicate systematic attempts to manipulate online discourse and manufacture the appearance of broad support, raising concerns about the integrity of electoral debate on TikTok. Yet the investigation also exposed hard ceilings on what can be proven, as TikTok's interface throttles scrolling after a few thousand comments; its anti-bot architecture blocks automation; and desktop collection cannot replicate the mobile experience that dominates Romanian usage. As a result, 25 % of comments remained unreachable, time-intensive manual capture was unavoidable, and moderation deletions may have gone unseen.

The Polish **cross-platform** case study reveals a large-scale, coordinated campaign on Polish X that amplified politically charged TikTok content, leveraging cross-platform sharing to flood the information space ahead of the 2025 presidential election. By mobilising over 600 synchronised accounts, operators exploited TikTok's share function to artificially boost engagement metrics, nudging its algorithm to further promote partisan content. The investigation demonstrates how design features, particularly automated cross-sharing, enable influence operations to obscure coordination while simultaneously maximising amplification. Findings highlight the emergence of a fragmented yet highly coordinated ecosystem, where domestic and foreign actors converge to distort political debate through ambient, high-volume posting strategies.

Taken together, the cases highlight overarching conclusions:

- **Measurement itself is brittle**: without transparent data access, audit reproducibility and comment-stream coverage quickly decay.
- **Algorithmic amplification hinges chiefly on subtle, implicit signals**, making its effects potent yet hard to contest through explicit user choice.

- **Coordinated manipulation tactics**: Both comment-section behaviour and cross-platform sharing illustrate how actors exploit TikTok to manufacture attention and simulate grassroots support.
- Research ceilings and blind spots: Technical barriers, such as throttled access to comments, moderation deletions, and inability to capture mobile experiences, impose hard limits on what researchers can detect or verify.
- Detecting manipulation requires local linguistic expertise and sustained platform cooperation; **Article 40 DSA pathways and current research APIs are too slow and restrictive to support real-time electoral safeguards**. Robust, longitudinal, user-centred and context-aware methodologies are therefore indispensable for credible oversight of TikTok's systemic risks.

6.3.2 Inauthentic Coordination Impact

Across the in-depth case studies, a coherent picture emerges of how coordinated actors weaponise visual media, platform loopholes, and generative AI to manipulate attention and belief while evading enforcement.

First, **all operations exploit asymmetries between paid advertising rules** and comparatively lack oversight of "organic" content. Facebook's strict ad policies on false news, gambling, and synthetic imagery are rendered as avoidable when identical material is posted as user-generated updates. The gambling network leveraged this gap most dramatically: 223 public groups published more than 10,000 posts per hour, yet faced little friction because they presented their promotions as community chatter rather than adverts. The papal-health hoax similarly slid past automated detection by burying the outbound link in the first comment and recycling near-duplicate collages that escaped Meta's existing fact-check labels.

Second, **generative AI acts as a force multiplier**. The gambling study documents a 13,242% surge in monthly posts after ChatGPT's public release, accompanied by a leap in visual sophistication, hyper-real lighting, fantasy motifs, and culturally localised narratives that amplified seven distinct persuasive drivers from luxury aspiration to fear-of-missing-out countdowns. In the Pope's case, an AI image still displaying the Grok xAI logo seeded hundreds of thousands of look-alike posts, illustrating how a single synthetic asset can spawn a self-replicating ecology of misinformation. For pro-Kremlin groups, AI assistance streamlined multilingual captioning and visual framing, enabling rapid production of heroic, militarised, and historically resonant portrayals of Vladimir Putin while ridiculing Western leaders.

Third, **coordination is both large-scale and structurally sophisticated**. All networks were surfaced by the vera.ai workflow through patterns such as high-frequency link sharing or synchronised posting, yet each developed distinct tactics: the gambling groups recruited users as peer promoters through referral incentives; the Pope-Francis campaign relied on centrally planted URLs funnelling traffic to a cluster of identically templated "breaking-news" blogs; the Putin fan pages synchronised static memes and immersive video keyframes to saturate feeds during politically salient windows, notably the 2024 U.S. election period.

Finally, the studies expose the **limits of current investigative and platform responses**. Meta's anti-spam and fact-checking systems failed to flag the majority of duplicate papal images; Meta's Content Library furnished only a fraction of the gambling posts because follower thresholds filtered out smaller groups;

and AI image analysis occasionally misidentified non-Western political figures. These blind spots underscore a regulatory dilemma: platform governance centred on content type, API restrictions, or English-language corpora cannot keep pace with multimodal, multilingual, AI-accelerated coordination. Sustained human-AI collaboration, unrestricted data access mechanisms, and impact-based policy triggers therefore emerge as critical prerequisites for mitigating the next wave of visual-first influence operations.

Across the audits, coordination cases, and project-wide recommendations, a coherent picture emerges. First, algorithmic amplification itself is a design-level risk: TikTok's "For You" feed personalises primarily on watch-time signals, funnelling users into ever-narrower loops and creating fertile ground for disinformation, yet reproducible audits show that small code or interface changes hamper oversight almost overnight. Second, the same attention architecture is readily weaponised by coordinated operators, who exploit platform loopholes and regulatory blind spots to circulate harmful narratives at scale while bypassing enforcement. These campaigns leverage generative-AI content, algorithmic curation, and cultural targeting to enhance reach and impact. Third, vera.ai tools demonstrate the promise and the limits of current responses: they can map cross-platform reach, detect coordinated link sharing, and flag suspicious bursts, but API lock-downs, the absence of qualitative indicators, and patchy access to recommender pipelines leave psychological influence, behaviour change, and the efficacy of counter-measures largely in the dark. Bridging that gap demands longitudinal, mobile-first audits; human-AI collaborative analysis; and, above all, meaningful data access under DSA Articles 40-42 so that amplification pathways – not just isolated pieces of content – can be traced, reproduced, and stress-tested. Mandating disclosure of recommender parameters and offering non-profiling feeds (Art. 38) would realign user agency, while impact frameworks that pair negative-effect metrics with evidence of mitigation would let projects like vera.ai track not only how far harmful narratives travel, but whether – and when – interventions truly blunt their force.

References

Annabell, T., Gorwa, R., Scharlach, R., van de Kerkhof, J., & Bertaglia, T. (2025). TikTok Search Recommendations: Governance and Research Challenges (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.08385.

Bastos, M. T., & Freitas, O. (2025). The Bannon doctrine: Network insurrectionism and democratic backsliding. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5186685.

Bastos, M., & Recuero, R. (2023). The Insurrectionist Playbook: Jair Bolsonaro and the National Congress of Brazil. Social Media + Society, 9(4), 20563051231211881.

Bauman, F., et al. (2025). Dynamics of Algorithmic Content Amplification on TikTok (arXiv:2503.20231v1). arXiv. https://doi.org/10.48550/arXiv.2503.20231.

Chaguri, M. M., & do Amaral, O. E. (2023). The Social Base of Bolsonarism: An Analysis of Authoritarianism in Politics. Latin American Perspectives, 50(1), 32–46. https://doi.org/10.1177/0094582X231152245.

Chakraborty, A., & Mattoni, A. (2023). Addressing corruption through visual tools in India: the case of three civil society initiatives and their Facebook pages. Visual Studies, 38(5), 803–816. https://doi.org/10.1080/1472586X.2023.2239759.

D'Onfro, J. (2016). Facebook's News Feed is 10 years old. This is how the site has changed. World Economic Forum. https://www.weforum.org/stories/2016/09/facebooks-news-feed-is-10-years-old-this-is-how-the-site-has-changed/.

Entrena-Serrano, Carlos & Degeling, Martin & Romano, Salvatore & Çetin, Raziye. (2025). TikTok's Research API: Problems Without Explanations. https://doi.org/10.48550/arXiv.2506.09746.

European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32022R2065.

Fenoll, V., Gonçalves, I., & Bene, M. (2024). Divisive issues, polarization, and users' reactions on Facebook: Comparing campaigning in Latin America. Politics and Governance, 12. https://doi.org/10.17645/pag.7957.

Giglietto, F., Marino, G., Mincigrucci, R., & Stanziano, A. (2023). A Workflow to Detect, Monitor, and Update Lists of Coordinated Social Media Accounts Across Time: The Case of the 2022 Italian Election. Social Media + Society, 9(3). https://doi.org/10.1177/20563051231196866

Giglietto, F., Righetti, N., & Rossi, L. (2020). CooRnet. Detect Coordinated Link Sharing Behavior on Social Media [Computer software]. https://github.com/fabiogiglietto/CooRnet.

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). 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.

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2021). CooRnet: An Integrated Approach to Surface Problematic Content, Malicious Actors, and Coordinated Networks. AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2021i0.12170.

Guimarães, S. S., Reis, J. C. S., Vasconcelos, M., & Benevenuto, F. (2021). Characterizing political bias and comments associated with news on Brazilian Facebook. Social Network Analysis and Mining, 11(1). https://doi.org/10.1007/s13278-021-00806-3.

Hao, K. (2021). The Facebook whistleblower says its algorithms are dangerous. Here's why. MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

Ibrahim, H., Jang, H. D., Aldahoul, N., Kaufman, A. R., Rahwan, T., & Zaki, Y. (2025). TikTok's recommendations skewed towards Republican content during the 2024 U.S. presidential race (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2501.17831.

Kemp, S. (2024). Digital 2024: Global Overview Report. DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2024-global-overview-report

Klug, D., Qin, Y., Evans, M., & Kaufman, G. (2021). Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. 13th ACM Web Science Conference 2021, 84–92. https://doi.org/10.1145/3447535.3462512.

Masood, M., Kannan, S., Liu, Z., Vasisht, D., & Gupta, I. (2025). Counting How the Seconds Count: Understanding Algorithm-User Interplay in TikTok via ML-driven Analysis of Video Content (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.20030.

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. New York: Data & Society Research Institute, 359, 1146-1151.

Mosnar, M., Skurla, A., Pecher, B., Tibensky, M., Jakubcik, J., Bindas, A., ... & Srba, I. (2025). Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings. arXiv preprint arXiv:2504.18140.

Mosseri, A. (2023). Instagram Ranking Explained. Instagram Blog. https://about.instagram.com/blog/spark/announcements/instagram-ranking-explained

Nimmo, B. (2020) The Breakout Scale: Measuring the Impact of Influence Operations. Washington, DC: The Brookings Institute.

Pamment, J. (2020). The EU's role in fighting disinformation: taking back the initiative.

Pamment, J., & Smith, V. (2022). Attributing information influence operations: Identifying those responsible for malicious behaviour online. NATO Strategic Communication Centre of Excellence.

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., & Kreslake, J. M. (2024). Beyond the margin of error: a systematic and replicable audit of the TikTok research API. Information, Communication & Society, 28(3), 452–470. https://doi.org/10.1080/1369118X.2024.2420032.

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). Available at https://computationalcommunication.org/ccr/article/view/120.

Peeters, S. (2023). Zeeschuimer (Version v1.4) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7525702.

Ridd, T. (2020) Active Measures. London: Profile Books.

Righetti, N., & Balluff, P. (2025). CooRTweet: A Generalized R Software for Coordinated Network Detection. Computational Communication Research, 7(1), 1. https://doi.org/10.5117/CCR2025.1.7.RIGH.

Recuero, R., Soares, F. B., Vinhas, O., Volcan, T., Hüttner, L. R. G., & Silva, V. (2022). Bolsonaro and the Far Right: How Disinformation About COVID-19 Circulates on Facebook in Brazil. International Journal of Communication Systems, 16(0), 24. https://ijoc.org/index.php/ijoc/article/view/17724.

Rodriguez, L., & Dimitrova, D. V. (2011). The levels of visual framing. Journal of Visual Literacy, 30(1), 48–65. https://doi.org/10.1080/23796529.2011.11674684.

Rogers, R. & Koronska, K. (2025). Post-Truth Spaces: Studying Authenticity and Influence on the Internet. International Journal of Communication.

Rogers, R. & Righetti, N. (2025). Coordinated Inauthentic Behaviour on Facebook? A Typology of Manufactured Attention. Platforms & Society.

Rossini, P., Baptista, É. A., Veiga de Oliveira, V., & Stromer-Galley, J. (2021). Digital media landscape in Brazil: Political (mis)information and participation on Facebook and WhatsApp. Journal of Quantitative Description: Digital Media , 1, 1–27. https://doi.org/10.51685/jqd.2021.015.

Ruediger, M., & Grassi, A. (2022). Polarization Presidentialism. How social media reshaped Brazilian politics: a case study on the 2018 elections. In Demokratie und Öffentlichkeit im 21. Jahrhundert–zur Macht des Digitalen (pp. 283-298). Nomos Verlagsgesellschaft mbH & Co. KG. 10.5771/9783748932741-283.

Thompson, C. (2020). YouTube's Plot to Silence Conspiracy Theories. Wired. https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories/

Vera, J. A., & Ghosh, S. (2025). "They've Over-Emphasized That One Search": Controlling Unwanted Content on TikTok's For You Page. https://doi.org/10.48550/ARXIV.2504.13895.

Vytautas, K. (2020). Deterrence: Proposing a more strategic approach to countering hybrid threats. The European Centre of Excellence for Countering Hybrid Threats, 9.

West, R., & Michie, S. (2020). A brief introduction to the COM-B Model of behaviour and the PRIME Theory of motivation [v1]. Qeios.

vera.ai

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.
The content of this document is © of the author(s) and respective referenced sources. For further information, visit veraai.eu.