Austrian Institute of Applied Telecommunications

January 2026 Report

Authors:
Julia Krickl
Natalie Trell
Lena Müller-Naendrup

Disinformation on Social Media Platforms

TGuard

Introduction

Disinformation is present throughout social media, and its forms range from organised campaigns during election periods and propaganda content in times of war and crisis, to pseudo-scientific, misleading or outright fabricated scientific content, and deceptively realistic AI-generated content and propaganda on polarising issues. Facebook, Instagram, YouTube and TikTok serve as central channels for everyday information, but they are increasingly also being used to spread such disinformation campaigns and narratives. Mis- and disinformation spreads not only through direct interactions, but particularly through algorithmic recommender systems, targeted advertising and group mechanisms, posing new challenges for authorities, civil society and researchers.

As designated Very Large Online Platforms (VLOPs) – platforms with more than 45 million users in the EU – these platforms are required under the Digital Services Act (Regulation 2022/2065) to take systematic action against illegal content and systemic risks such as disinformation. Key elements include notice-and-action mechanisms, transparency obligations and cooperation with Trusted Flaggers, which are recognised entities with particular expertise whose reports regarding illegal content must be given priority by platforms. Telegram is also included in this study, though it falls outside the VLOP framework and is therefore treated separately, as outlined in the approach section below.

The TGuard[1] research project examined the extent of disinformation on the above-mentioned social media platforms and the measures they have taken to curb it. The central question was how effective the moderation and regulatory measures deployed against disinformation are in the context of the DSA, and where structural gaps exist between the platforms' self-reporting and the observations of independent investigations.

Approach

The analysis follows a cross-platform comparative approach to situate the extent and character of disinformation in the digital space, drawing on the following sources:

TRANSPARENCY REPORTS

Most recent semi-annual transparency reports submitted by the platforms under the Code of Practice on Disinformation (CoPD)[2], self-reported by platforms.

DATA

from the DSA Transparency Database[3] for 2025, containing the „Statements of Reasons" (SoRs) for moderation decisions.

TOPIC-SPECIFIC STUDIES

Investigations by independent organisations such as the European Fact-Checking Standards Network (EFCSN), Maldita, Science Feedback, AI Forensics and Democracy Reporting International (DRI).

[1] https://www.tguard.at/

[2] Since February 2025 Conduct on Misinformation. Analysis: CoPD-Report No. 6 for Meta, No. 5 for TikTok and Google/YouTube, reporting period up to the end of 2024 or mid-2025. The reports are available here: https://disinfocode.eu/reports

[3] https://transparency.dsa.ec.europa.eu/

Where platform figures and external investigations diverge, both perspectives are made explicit. The analysis aims to draw a systematically comparable picture of platform-specific practices and their limits, as well as to illustrate key disinformation narratives across platforms. Building on this, three dimensions were examined for Meta (Facebook and Instagram), YouTube and TikTok respectively:

1. the applicable platform policies on disinformation, political advertising, AI content and problematic accounts;

2. the detection procedures deployed by the platforms and the countermeasures derived from them;

3. the effectiveness of these measures based on independent investigations and documented case examples.

Telegram is treated separately in an excursus, as the platform is not classified as a VLOP under the DSA and does not participate in the CoPD, meaning key comparative data are missing. The excursus therefore follows an alternative approach and relies more heavily on the platform's development, usage dynamics and findings from academic and civil-society analyses.

Cross-Platform Data

DSA-Transparency Database: Statements of Reason 2025

The DSA Transparency Database documents every moderation measure taken by the platforms, together with a justification ("Statements of Reasons", SoRs) and the type of measure. There is no dedicated category for disinformation; relevant violations may fall under various categories, in particular "Negative effects on civic discourse or elections" and "Illegal or harmful speech".

Regarding the types of violation: for TikTok, almost 50 % of all SoRs fell under the categories relevant to disinformation. The Meta platforms, by contrast, recorded no SoRs under "Negative effects on civic discourse or elections"; the largest category besides "Scope of platform" was "Scams and/or fraud". YouTube did report relevant cases, but proportionally these accounted for less than 1 % of total SoRs.

The types of measure show differing patterns: TikTok responded predominantly with visibility restrictions (990 million cases; content was removed in just under 60 % of cases), Facebook primarily with account restrictions (359 million cases, of which 258 million were suspensions), YouTube with content removals (95,8 million) and Instagram with account terminations

Platforms	Total SoRs 2025	of which: negative effects on civic discourse	of which: illegal or harmful speech
TikTok	1,002,728,224	45,586,463	434,875,107
Facebook	453,069,643	-	15,821,557
Instagram	126,151,906	-	2,812,557
YouTube	104,659,285	178,267	56,478

Table 1: SoRs 2025 by platform (consolidated according to categories relevant to disinformation). Source: DSA-Transparency Database, own presentation.

(104,6 million). Notably, most of the moderation measures stemmed from voluntary platform initiatives. Reports under Art. 16 DSA, which requires publicly available and easily accessible reporting procedures for users, and reports from Trusted Flaggers played a minor role. User reports accounted for only a small share of total measures, ranging from just under 150,000 (TikTok) to fewer than 500,000 (YouTube), while Trusted Flagger reports amounted to fewer than 500 cases per platform.

CoPD-Reports: Moderation Measures H1 / 2025

The CoPD is a voluntary self-regulatory framework, originally introduced in 2018 and revised in 2022. Signatories – including the platforms analysed here, except for Telegram – commit to submitting detailed transparency reports every six months along defined „commitments", comprising both quantitative data and qualitative descriptions. Of particular relevance to this analysis are the commitments on „Scrutiny of Ad Placements" (restricting the monetisation of disinformation), „Integrity of Services" (countering coordinated manipulation) and „Empowering Users" (contextualising misleading content). The CoPD reports provide platform-specific figures on moderation measures according to each platform's own policy categories. The **table 2** below consolidates the key figures for the first half of 2025 by topic, and content or accounts removed.

The absolute figures are only comparable to a limited extent, since the platforms define the policy areas differently and record content at different levels (individual videos vs. accounts). More informative is the following pattern: TikTok provides the most extensive and detailed data, YouTube the most transparent breakdowns by area, and Meta the lowest granularity. For the categories relevant to disinformation, Meta didn't report information for Facebook and Instagram separately in some cases.

A further aspect addressed in the CoPD reports is networks of organised, covert influence, or Coordinated Inauthentic Behaviour (CIB), used to spread political disinformation and to influence public opinion. The documentation of action taken against such networks differs between platforms: some rely on quantified data, some on more detailed qualitative descriptions. During the reporting period, TikTok removed at least 16 CIB networks with a total of 31,270 accounts and just under 10.5 million followers. Meta reported the removal of several CIB networks, including one focused on Romania (658 Facebook accounts, 14 pages, two Instagram accounts). YouTube, in turn, provided exclusively quantitative data without any substantive description: 177,272 fake channels were deleted EU-wide (838 in Austria), and the platform additionally identified 4,284 CIB cases at EU level and 71 in Austria.

Platform	Misinformation		Synthetic Media		Fake-Accounts / Channels	
	EU	AT	EU	ET	EU	AT
TikTok	207,646	2,709	24,213	585	19,225,031	285,956
Meta	83,000+	1,000+	-	-	~1B (globally)	-
YouTube	4,060	62	1,776	29	177,272	838

Table 2: Moderation measures according to CoPD transparency reports for the first half of 2025. Sources: TikTok 2025, Meta 2025, Google/YouTube 2024/2025, own presentation.

Example Narratives

Beyond the quantitative data, case studies show which disinformation narratives circulate on the platforms and how far their reach extends despite moderation:

- Election- and democracy-related narratives: In the context of the 2024 European elections, social media platforms were a central target for various disinformation campaigns and narratives seeking to influence democratic processes. Algorithmic recommender systems played a central role in the spread of political content (Fundación Maldita, 2024).

- AI-assisted war narratives: In an ÖIAT investigation, 18 of 21 identified AI videos related directly to the conflicts in Israel and between Israel and Iran – often featuring entirely artificially generated civilians or soldiers. An AI-generated, purported missile strike on Tel Aviv was only removed after a Trusted Flagger report and had, by that point, already reached more than 200,000 views (Strasser & Krickl, 2025).

- Health and climate disinformation: A Facebook video claiming that a CO_2 concentration below 0.03 % would destroy all life on Earth remained online despite a fact-check, with more than 250,000 views. On YouTube, a video with the false claim that the EU's ban on combustion-engine cars from 2035 had been repealed reached 500,000 views without any platform response (Fundación Maldita, 2024).

Platform-Specific Analysis

Meta (Facebook and Instagram)

Policies

Meta prohibits the dissemination of false information that has the potential to cause physical harm or to undermine democratic processes. Content that is demonstrably false but poses no immediate danger is not removed, but supplemented with factual counter-statements. Since October 2025, political, election-related and social-issue advertising has been entirely prohibited on Meta in the EU. AI-generated or manipulated content has been labelled more extensively since May 2024, particularly in the case of deepfakes with societal relevance.

Detection and Measures

Meta relies on a combination of automated systems, human review and cooperation with external fact-checking partners. As of 2024, Meta works with 29 organisations in the EU that review content in 23 languages. If a post is classified as false, it receives a visible label with an explanation and reduced visibility in the algorithm. For the 2024 European elections, Meta set up "Voter Information Units", which, according to its own figures, were viewed over 41 million times on Facebook and 58 million times on Instagram (Meta, 2024).

In January 2025, Meta announced that it would end the third-party fact-checking programme in the US and switch to a Community Notes model. In Europe, the fact-checking programme is being continued for now due to DSA obligations. The new direction – politically motivated and shaped by the climate in the US – has alarmed fact-checking organisations, particularly with regard to a possible further withdrawal from CoPD commitments (Schulz-Tomančok, 2025).

Effectiveness

Despite extensive policies, the effectiveness of Meta's

measures remains contested. A Europe-wide study by the fact-checking organisation Maldita.es (conducted within the framework of the EFCSN) analysed 1,321 disinformation posts reported to platforms and produced two key findings regarding Meta (Fundación Maldita 2024)

- High response rate, low removal rate: On Facebook, measures were taken in around 89 % of cases (Instagram: 71 %), but the content was removed in only 12 % (Facebook) and 14.6 % (Instagram) of cases. The majority of measures were limited to attaching a fact-check label.

- Thematic inconsistency: For specific topics, the response rate was significantly lower – only 33 % for climate-related disinformation on Instagram, and just over 50 % for content on the Middle East conflict and on national and regional topics.

Research by AI Forensics reveals considerable weaknesses in the enforcement of the advertising policies: around 60 % of the moderated political ads in the Meta Ad Library had already violated applicable platform policies, while automated systems detected fewer than 5 % of the problematic ads. One documented pro-Russian propaganda campaign reached more than 38 million users in France, Germany, Italy and Poland (Bouchaud et al., 2024). Science Feedback additionally identified at least 21 Facebook accounts in 2024 with links to sanctioned Russian or Belarusian actors; by the end of the year, nine of these profiles remained active despite being reported, and 13 newly added accounts remained unmoderated (Terroille et al., 2024).

YouTube

Policies

YouTube prohibits content that deliberately spreads false information with "significant potential for harm", in particular regarding elections, censuses and health-related topics. Other false or misleading content remains online but is restricted in its reach and provided with context. Since March 2024, creators have been required to openly disclose realistically altered or synthetic content; technical detection mechanisms such as SynthID complement this through invisible watermarks.

Detection and Measures

YouTube combines automated AI-assisted classification, human moderation and cooperation with certified fact-checking organisations. Early-detection systems watch for unusual upload or interaction patterns that indicate coordinated campaigns. Countermeasures include removal, reach reduction, information panels and the promotion of "authoritative sources".

Effectiveness

In its CoPD report for the first half of 2025, YouTube reported comparatively low absolute figures: 4,060 misinformation removals EU-wide, 62 in Austria. For coordinated content amplification, there were 4,284 cases EU-wide. Whether these figures reflect the actual burden or point to systematic under-recording is difficult to assess without external data access.

A sample investigation by Science Feedback found that, in an assessment of 940 YouTube videos classified as problematic, an information panel was displayed in only around 37 % of cases – even though YouTube prominently communicates this instrument as a measure against disinformation (Samson, 2024). As the case examples in section 2.3 show, high-reach false information can remain online without any platform response – as with the video viewed 500,000 times about the supposedly repealed EU combustion-engine ban. Another YouTube reel with the false claim that COVID-19 vaccinations trigger autoimmune diseases was viewed more than 300,000 times in Poland alone (Fundación Maldita, 2024).

TikTok

Policies

TikTok's Community Guidelines prohibit "harmful misinformation", including content that can endanger public safety, health or democratic processes. A far-reaching ban applies to political advertising: paid political ads are not permitted on TikTok. Politicians, governments and political parties lose access to advertising functions on verified accounts. Synthetic media that realistically depict real people or scenes must be labelled as such.

Detection and Measures

TikTok relies heavily on automated moderation: a large share of moderation decisions is made semi- or fully automatically. In addition, the platform cooperates with the Global Fact-Checking Program (GFCP) and certified fact-checkers. In response to identified disinformation, videos can be removed, blocked from the For You feed, or provided with warning labels and information banners. In the case of coordinated campaigns, entire networks are removed.

Effectiveness

TikTok provided the most extensive figures in the official reports: around 207,000 misinformation videos and more than 19 million fake accounts were removed in the first half of 2025 EU-wide. One focus is on structural interventions against coordinated manipulation: 16 CIB networks with around 10,5 million followers were removed in the same period.

The actual implementation in everyday practice, however, shows considerable gaps. An investigation by CORRECTIV and other organisations into the 2025 German federal election found that, of 201 reported disinformation videos, only 12 – around 6 % – were actually removed from the feed. Similar patterns emerge from DRI's investigations into the 2025 Romanian presidential election, in which TikTok addressed 323 accounts identified as problematic only late in the electoral process. Independent studies have also repeatedly documented circumventions of the ban on political advertising: researchers found 100 political ads on TikTok Germany within a single investigation (Hohlfeld et al., 2024; Kimberly and Scherndl, 2025).

12 %
removed

Only 12 % of the disinformation content reported on **Facebook** is actually taken down. There is a response rate of 89 % of it, but usually only in the form of a label.

37 %
labelled

Of 940 items classified as problematic on **YouTube**, 37 % received an information panel.

100
political ads

Despite the ban on such advertising, researchers found 100 political ads in the course of one study on **TikTok**.

Excursus: Telegram

A Special Position within the Regulatory Framework

Telegram occupies a special position: the platform states that it has fewer than 45 million monthly active users in the EU and is therefore not classified as a VLOP under the DSA.

This means Telegram is under no legal obligation to identify or assess systemic risks such as disinformation, or to document measures to curb them. Nor does it participate voluntarily in the CoPD. Moreover, private messaging services are exempt from central DSA obligations – a categorisation that Telegram claims for itself despite its public channel function.

Their current terms of service contain no explicit indication that disinformation is prohibited. According to Telegram, content that calls for violence or terrorist propaganda is moderated, and content relating to the sexual abuse of minors is removed. Moderation thus concentrates on a narrow set of content categories and follows a primarily reactive logic. Enforcement of the terms of service is also limited to publicly designated groups and channels. Administrators alone decide whether a channel counts as public, meaning high-reach communication spaces can formally be run as private.

Development into a Platform with Disinformation

Telegram usage doubled during the COVID-19 pandemic from 2020 onwards, and much of the far-right and conspiracy-ideology scene concentrated its communication on the platform. As early as 2020, around 96 % of the most influential actors in German-language right-wing extremism operated a channel on Telegram, according to the Institute for Democracy and Civil Society (Donner et al., 2025). With Russian's war of aggression in 2022, Telegram also became a distribution channel for geopolitical disinformation: after the YouTube ban of the Russian state broadcaster RT DE, it moved to Telegram, where the pro-Russian influencer Alina Lipp reached an average of 75,000 views per message in December 2022 (Holnburger, 2023).

Studies show that disinformation on Telegram becomes effective not primarily through broad mass distribution, but through a small but highly engaged user base. Posts from reputable sources with links to external sources do appear up to three times more often than misleading ones, but channels with misleading content are up to 18 % more active. The higher the share of misleading sources, the livelier the internal exchange. This creates viral effects that do not depend on algorithmic amplification: virality arises primarily from coordinated posting, mutual cross-linking and the intensive activity of small but well-connected groups (Herasimenka et al., 2023).

Assessment

Telegram combines the public reach and networking logic of social media with the regulatory self-image of a private messenger. This combination allows the platform to exploit the infrastructural advantages of social media without assuming the corresponding regulatory responsibility. The European Commission's recent designation of WhatsApp as a VLOP on 26th January 2026 – expressly on account of its public-facing "Channel" feature – underscores the problem. Telegram exhibits functionally comparable, and in some respects more far-reaching, public distribution mechanisms, yet so far remains below the VLOP threshold (Digital Strategy EC, 2026).

Conclusion & Recommendations

Disinformation is present on all major online platforms and gains particular relevance in politically and socially sensitive contexts. It spreads not only through individual viral events, but through longer-term dynamics that influence public debate, the formation of political opinion and trust in democratic institutions. Repeated and interconnected disinformation narratives contribute to the erosion of shared points of reference and deepen existing polarisation.

The examples analysed in this dossier show that social media platforms are not neutral distribution instruments; through their technical, economic and communicative platform logics, they substantially shape which content gains visibility. Algorithmic recommender systems, group-based communication spaces with broadcasting functions and increasingly AI-assisted content, all amplify the presence and spread of polarising, misleading or manipulated content.

What the DSA achieves and where it reaches its limits

The DSA, the DSA Transparency Database and the CoPD create a central framework for capturing disinformation comparably across platforms and holding platforms more accountable. Systemic risk assessments and the obligation to mitigate risk help counter coordinated campaigns in the political sphere. However, measures such as the CoPD reports remain voluntary, and are thus dependent on the cooperation of the major social media platforms.

The example of Telegram, however, makes clear that moderation reaches its limits where regulatory obligations do not apply. The designation of WhatsApp as a VLOP in January 2026 – expressly on account of its public channel function – illustrates a central weakness of current platform regulation: where public impact does not coincide with regulatory classification, disinformation dynamics can spread and take hold.

Limits of Self-Reporting

The effectiveness of internal platform measures remains only partly assessable without external oversight, because essential data and assessment logic continue to reside with the platforms themselves, and reporting formats rely mainly on self-disclosure. The CoPD functions largely as an “à la carte” instrument: signatories choose for themselves which commitments to implement for their service. This approach undermines effectiveness and increases the need for regulatory and civil-society scrutiny.

This is even more problematic given that in recent years platforms have partly scaled back measures or implemented them only selectively. Meta‘s US switch from third-party fact-checking to Community Notes, and its political framing, represent a visible strategic shift whose effects in Europe will only become apparent in the future.

Key Recommendations

- → Align regulatory classification with actual public impact: Platforms with public distribution mechanisms should not be able to escape VLOP obligations by self-positioning as a “private messenger”.
- → Complement self-reporting with independent oversight and data access: Fact-checking organisations, civil-society monitoring and independent research are indispensable complements to the platforms‘ reports. A prerequisite for this, however, is reliable access to platform data – for example, via the data access rights for “vetted researchers” provided for in Art. 40 DSA.
- → Make CoPD commitments more binding: The voluntary basis and the possibility of selectively choosing commitments undermine the effectiveness of the instrument and should be replaced by binding standards.

→ Require cross-topic consistency: The lower response rate for climate disinformation, migration and geopolitical topics shows that moderation measures must be effective not only for mainstream political content.

Only through the interplay of clear obligations, independent observation and consistent cross-platform classification can structural disinformation dynamics be addressed sustainably.

Note on Sources and Further Information

This short version is based on the detailed dossier "Disinformation on Social Media Platforms" (ÖIAT, January 2026), researched and published as part of the TGuard project (K-Pass / KIRAS). The full dossier contains detailed data tables, methodological explanations and a complete bibliography of around 60 sources, including reports by the European Commission, the platforms' CoPD transparency reports, academic publications and investigations by independent organisations (EFCSN, Maldita, Science Feedback, AI Forensics, Democracy Reporting International, CeMAS, among others).

The full dossier is available via **research.oiat.at**.

Sources

Bouchaud, Paul, Marc Faddoul, und Raziye Buse Çetin. 2024. No Embargo in Sight. Meta Lets pro-Russia Propaganda Ads Flood the EU. AI Forensics. https://aiforensics.org/work/meta-political-ads.

Digital Strategy EC. 2026. „Commission designates WhatsApp as Very Large Online Platform under the Digital Services Act". Januar 26. https://digital-strategy.ec.europa.eu/en/news/commission-designates-whatsapp-very-large-online-platform-under-digital-services-act.

Donner, Christian, Wyn Brodersen, Maik Fielitz, und Michael Schmidt. 2025. „Drei Jahre MATR-Monitoring auf Telegram: Die Rolle der Plattform im rechtsalternativen Online-Aktivismus". Machine Against the Rage, Nr. Nr 8. https://doi.org/10.58668/matr/08.1.

Fundación Maldita. 2024. Platform Response to Disinformation during the EU Election 2024. Fundación Maldita. https://maldita.es/nosotros/20240611/platforms-disinformation-response-elections-eu/.

Herasimenka, Aliaksandr, Jonathan Bright, Aleksi Knuutila, und Philip N. Howard. 2023. „Misinformation and professional news on largely unmoderated platforms: the case of telegram". Journal of Information Technology & Politics 20 (2): 198–212. https://doi.org/10.1080/19331681.2022.2076272.

Hohlfeld, Alexander, Anna Semenova, Martin Degeling, Greta Hess, und Kathy Meßmer. 2024. Tik-Tok, DSA O'Clock? auditing TikTok. Interface. https://tiktok-audit.com/blog/2024/Tik-Tok-oclock/.

Holnburger, Josef. 2023. Chronologie einer Radikalisierung: Wie Telegram zur wichtigsten Plattform für Verschwörungsideologien und Rechtsextremismus wurde. CeMAS - Center für Monitoring, Analyse und Strategie. https://report.cemas.io/telegram/.

Kimberly, Nicolaus, und Gabriele Scherndl. 2025. „Melden zwecklos: Tiktok ignoriert Desinformation zur Bundestagswahl". correctiv.org, Februar 21. https://correctiv.org/faktencheck/hintergrund/2025/02/21/melden-zwecklos-tiktok-ignoriert-desinformation-zur-bundestagswahl/.

Meta. 2024. Code of Practice on Disinformation -Report of Metafor the Period 01 January to 30 June 2024. Nr. 4. Code of Conduct on Disinformation. Transparency Center. https://disinfocode.eu/reports.

Samson, Charlotte. 2024. „Use of information panels on YouTube videos containing misinformation – a 'quick check' study -". Science Feedback, März 19. https://science.feedback.org/information-panels-youtube-videos-containing-disinformation-quick-check-study/.

Schulz-Tomančok. 2025. „Meta gegen Europa: Zensur oder Pflicht zum Fakten-Check?" https://www.oeaw.ac.at/news/meta-vs-eu-der-kampf-um-faktenchecks.

Strasser, Iris, und Julia Krickl. 2025. Falsch- und Desinformation auf Social Media. Meldung problematischer Inhalte auf TikTok und Instagram als Trusted Flagger. ÖIAT. https://research.oiat.at/fileadmin/Research/Dokumente/Falsch-Desinfo_SocialMedia_%C3%96IAT_15072025.pdf.

Terroille, Charles, Kateryna Chystoforova, und Saman Nazari. 2024. Sanctioned but Thriving: How Online Platforms Fail To Address the Widespread Presence of Entities Under EU Sanctions. Science Feedback. https://science.feedback.org/sanctioned-but-thriving-how-online-platforms-fail-to-address-the-widespread-presence-of-entities-under-eu-sanctions/.

Imprint

Researched as part of the cooperative R&D project TGuard and funded under the security research funding programme.

Bundesministerium
Finanzen

FFG
Forschung wirkt.

K-PASS

Research conducted by:

Authors: Julia Krickl MA | Natalie Trell MA | Lena Müller-Naendrup MSc

ÖIAT Research
Austrian Institute for Applied Telecommunications
www.research.oiat.at
Ungargasse 64-66/3/404
1030 Vienna

January 2026