March 2026

Input on recurrent and prominent systemic risks in the EU and on measures for their mitigation

Maldita.es is a non-profit foundation based in Spain that builds public trust and protects information integrity through journalism, education, technology, research and policy action. Our work is underscored by specialised teams, cutting edge technological tools, and an extensive community of citizens who collaborate with us in the battle against disinformation.

Contact point: policy@maldita.es

QUESTION 1 (ON SYSTEMIC RISKS)

The risk presented by **AI-generated disinformation** has increased exponentially during the last period. With the improvement of the quality of the outputs produced by AI models, these tools are being used to produce realistic content linked to current world events, displacing legitimate coverage and shaping public opinion. Despite community guidelines aimed at this type of content, several investigations published by Fundación Maldita.es reflect this trend in various VLOPs:

- 550 TikTok profiles posted over 5,800 AI-generated videos of massive protests related to real world events, such as the capture of Maduro in Venezuela, civil protests in Iran, or general elections in Chile, fuelled by monetization incentives. Some of these videos amassed up to 4 million views while the profiles reached thousands of new followers through these videos portraying false scenes that affect civic discourse.

- Networks of YouTube channels mass produce and post videos containing synthetic voices and supporting false claims on AI-generated media. A collected sample of 49 channels shared disinformation about confrontations between European politicians and Spanish politicians in the European Parliament that had never taken place, accumulating 32.2 million views.

- Realistic videos generated with artificial intelligence have also increased in relation to the war in Ukraine to promote disinformation narratives. These are disseminated in various online platforms, showing wounded soldiers surrendering to the Russian army, being forced to go to the front, or expressing regret for having enlisted as volunteers. For instance, one AI-generated video of an Ukrainian soldier crying because of being forced to go to war was posted in at least 13 different languages as real footage. One of the posts reached 710,000 views on X in two days.

Specifically on TikTok, Fundación Maldita.es has looked into the risk of **minors becoming victims of sexualization** online. Profiles either compiling videos of real minors or of AI-generated media portraying underaged-looking girls in suggestive positions collected hundreds of comments of sexual nature and remained accessible despite our notifications. Moreover, Telegram accounts where CSAM is sold were promoted in the comment section of these videos.

VLOPs such as Facebook or TikTok are used to **disseminate illegal content**. For instance, Fundación Maldita.es detected 1,075 Facebook pages, part of a campaign of impersonation of public transportation services of over 740 cities around the world used for phishing scams. On TikTok, over 100 profiles promoted apartments to squat in Spain, which can be considered a crime under the Spanish Penal Code.

QUESTION 2 (ON RISK FACTORS)

The investigations conducted by Fundación Maldita.es are evidence of various risk factors linked to the systems of VLOPs that are allowing and even incentivizing the dissemination of harmful and/or illegal content.

Fundación Maldita.es has reported deceptive AI-generated media which is not essentially motivated by social or political stances but whose goal is to generate revenue. Accounts posting AI-generated videos of protests worldwide were incentivized to produce disinformation by TikTok's **monetization program**. Owners of the profiles claimed disinformation on current events is the most efficient way to generate engagement and reach 10,000 followers needed to join the "Creator Rewards Program". These harmful videos were a fast path to enable content monetization in the platform but also to build an engaged audience and proceed to **sell the profile**, not allowed under the Terms of Service.

In relation to financial incentives, profiles disseminating sexualizing content of minors are making use of TikTok's **subscription system** (from which the owner of the profile and the platform take a share of the profits) to share 'exclusive' harmful content. Posting this content where minors are sexually objectified is a way to attract users to subscribe and generate revenue. While Open AI's Sora was used to generate a fraction of AI-generated media of minors, TikTok AI Alive watermark was visible in several videos. This reflects **the integration in VLOPs' systems of generative AI tools** without adequate safeguards that provokes damage to minors or to women, as in the case of xAI's Grok.

The evidence for this investigation was collected almost exclusively through TikTok's **algorithmic recommendation** in its "For You" section, which became a feed of sexualizing videos of minors. Moreover, TikTok's "You May Like" search suggestions directed to more of these videos.

Algorithmic amplification was also tested by Fundación Maldita.es and AI Forensics in relation to climate disinformation in YouTube and TikTok. Specifically, the performance of

posts containing debunked claims about 2024 floods in Valencia and shared during the month that followed the catastrophe was assessed. A significant overall reach on both platforms indicates **platform-level amplification**. Each dis/misinformation video about the event on YouTube had an average of 21,250 views which is almost four times the platform average. On TikTok these videos had 32,294 views on average.

VLOP's advertisement systems are heavily exploited for inorganic amplification of harmful/illegal content. For instance, Fundación Maldita.es reported on two Meta pages which spent almost 40,000 euros on ads targeting political parties in Spain, including the promotion of disinformation narratives. Scammers keep relying on paid ads to direct users to fraudulent websites. Facebook pages pretending to be public transportation services in 60 different countries launched more than 9,000 ads to increase their reach. Many brands are also victims of impersonation in ads on TikTok and Instagram: 37 advertisers posted 18,000 ads directing to deceiving sales. This has been possible to evaluate as minimum data is provided through their ad repositories. Meanwhile, other VLOPs lack adequate information or interfaces for their mandatory ad repositories preventing independent assessment of potential risks.

Gaps in **content moderation and repeated offenders considerations** are very much linked to this risk factor. All scammer advertisers collected in TikTok and Instagram had at least one ad moderated, still 79 percent of them were still active on the platforms. In the case of the public transport scam, all of the pages were accessible even if 55 percent of them had been previously moderated. Even if **reporting mechanisms** were to be used, the results are similar: 93 percent of profiles impersonating public transport flagged to Facebook by Fundación Maldita.es were still active, even if these pages were systematically being repurposed to target other cities.

Moderation failures take place despite the adequacy of community guidelines. While TikTok does not allow "accounts focused on AI images of youth in sexualized poses", 93 percent of those reported by Fundación Maldita.es under that specific ground were still running. Even if flagging specific videos, TikTok found no violation of their policies in 76 percent of the cases. This could be caused by **overreliance on automated systems** incapable of evaluating sexualizing content even if obvious to the human eye.

QUESTION 3 (ON MITIGATION MEASURES)

Contextual information alongside posts that contain false or misleading claims: Rather than removing content, value is found in platforms' adding explanations that provide users with relevant facts and sources. This approach respects freedom of speech while empowering users to make informed decisions about the claims they encounter. Importantly, contextual information is most effective when it is highly specific to the post in question, directly addressing the claims made rather than offering vague warnings.

- Fact-checking partnerships: One of the most widely accepted ways to provide contextual information is through collaborations with independent fact-checking organizations. These partnerships allow platforms to integrate conclusions reached through transparent methodologies and supported by local expertise. Evidence suggests this approach reduces the dissemination of mis/disinformation and influences user behavior. A study published in December 2025, led by Julia Cagé at Sciences Po in Paris, found that when content is labeled as "false", the users who shared it are twice as likely to delete their post and are also less likely to spread misinformation in the future. This indicates that credible, expert-driven verification can significantly discourage the circulation of misleading claims.

- Voluntary crowdsourcing: Another approach relies on users participation, such as the Community Notes system implemented on several platforms. In this model, users can propose contextual notes to posts, and the visibility of those notes is determined by agreement among contributors rather than factual accuracy. While this method offers scalability and a decentralized mechanism, it has clear limitations. Many notes are never displayed, the system does not necessarily rely on expert evaluation, and its effectiveness decreases during moments of crisis when rapid and authoritative information is most needed.

Evidence from platform implementation illustrates these limitations. Meta's chief information security officer reported that 900 Community Notes became visible in the first six months after the system was rolled out in the United States. This figure appears especially modest when compared with the scale of traditional fact-checking interventions. Over a similar period, Meta applied fact-checking labels to around 35 million Facebook posts in the European Union alone.. Moreover, the system as it stands is already showing signs of fatigue, which raises concerns over long term sustainability.

A mixed approach appears to be the most effective path forward. Platforms can combine the scalability and participatory nature of Community Notes with the methodological rigor and credibility of professional fact-checking organizations. By integrating both mechanisms, online platforms can expand coverage while maintaining high standards of accuracy, ultimately strengthening efforts to curb the spread of disinformation without undermining open online discourse.

QUESTION 4 (ON ADDITIONAL INFORMATION)

Cross-platform impact of systemic risks: Most cases illustrating systemic risks in different very large online platforms (VLOPs) show that these risks are often not confined to a single service but instead spill over into other parts of the online services ecosystem. Harmful content or behaviors frequently operate across multiple platforms, meaning that insufficient mitigation in one service can facilitate harm elsewhere.

One illustrative case concerns inadequate moderation of comments under videos that sexualize minors on TikTok. In some instances, these comment sections have been used to promote external channels that direct users to Telegram, where child sexual abuse material (CSAM) may be exchanged. When such comments remain visible and unaddressed, they function as a gateway, allowing harmful networks to recruit or redirect users from a mainstream platform to less moderated environments.

A related challenge arises from the current **lack of interoperability** between mitigation measures implemented by different platforms. Even when one platform introduces safeguards to identify or contextualize harmful content, these are not displayed when the same content is redistributed elsewhere. A common example involves AI-generated content that is labeled as such on a platform. Once downloaded and reposted on another platform, the label may disappear, lacking contextual information that signals its synthetic nature, and increasing the risk that it turns into misinformation.

Unmitigated risks on non-VLOPs: Fundación Maldita.es continues to collect evidence on the dangerous practices taking place on online platforms not designated as very large online platforms and thus, not legally obliged to evaluate and mitigate risks. This is the case with Telegram. For instance, an investigation by Maldita reported on channels being used as a dark market for non-registered weapons.