April 2025

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.

Our mission is to provide all actors affected, from legislators and digital platforms to journalists, citizens & educators, with tools, capacities, and evidence-based research so that they can make informed decisions, and together we can foster a more resilient, accessible, and trustworthy media & information ecosystem.

Contact point: policy@maldita.es

---

1. Introduction

Fundación Maldita.es has gathered first-hand evidence from both Spain and the broader European Union on systemic risks identified across various online platforms. While the majority of the research focuses on the spread of harmful disinformation, it also includes documented instances of illegal content dissemination.

The contribution assesses the effectiveness of commonly used visible measures to counter misinformation and disinformation, such as fact-checking labels, informational panels, or removal by certain platforms.

Additionally, it examines how advertising systems are facilitating the wide promotion of harmful content, as well as the deceptive role of X's blue check system.

2. Identified Recurring Risks

2.1. Harmful disinformation resulting in real harm to Public Security, Civic Discourse, and Fundamental Rights (Article 34(1)c)

Unaddressed systemic risks inevitably surface and cause the greatest harm when unexpected events occur. This presents a significant challenge for all very large online platforms in managing harmful disinformation, which can severely impact citizens' security and well-being, as well as trigger widespread civic unrest.

Fundación Maldita.es has monitored how widespread and unaddressed misinformation narratives are adapted to specific situations, becoming even more dangerous when fueled by anger, frustration, or desperation. This is evident during major natural disasters such as the deadly floods in eastern Spain in 2024, when disinformation about artificial climate manipulation and the removal of dams and reservoirs, repeatedly debunked by Maldita.es,

resurfaced in public discourse during the catastrophe all but ensuring that even first responders had to operate sometimes in a hostile, polarized environment.

Even when an event is perceived as geographically limited, disinformation knows no borders. In the aftermath of the floods, Maldita tracked how debunked hoaxes and conspiracy theories spread beyond Spain, reaching audiences in Europe, Latin America, the United States, and India, where they were weaponized to serve ideological or political agendas.

A similar pattern emerged after the stabbing of an 11-year-old boy in a Spanish small town, as disinformation and unverified claims about the perpetrator's origins circulated not only in Spain but also in multiple languages across European countries. While, in this case, public institutions were able to swiftly counter the falsehoods, such narratives found fertile ground for dissemination in well-established anti-immigration networks, openly inciting real-world violence, just as seen after the stabbing in Southport (England) in July 2024.

Those are just two examples in a long list of instances in which the major platforms' failure at mitigating the risk of unaddressed disinformation in their services resulted in clear real-life harms to not only public security and harms, but also the fundamental rights of minorities.

2.2. Harmful Disinformation in Electoral Processes (Article 34(1)c)

Elections and other democratic processes are well-known targets for disinformation, with a significant increase in misleading and manipulated content online during these periods. Unlike unexpected threats referenced previously, these events are planned months in advance, theoretically allowing for better preparation to prevent and respond to harmful content, many of which have been repeatedly observed in the past.

Fundación Maldita.es has identified disinformation narratives spreading across online platforms that target democratic processes. Through its daily monitoring and debunking efforts. Moreover, it has collaborated in EU-wide initiatives such as Elections24Check, where a database that gathered and categorized fact-checked information for European countries and citizens ahead of the 2024 European Parliament Elections was built, allowing for cross-border comparison.

In Spain, though similar patterns were observed in other countries, the most prominent disinformation claims on social media during the EU Elections targeted the electoral process itself. These narratives misled voters by encouraging incorrect voting practices, alleging an "information blackout," or suggesting manipulation of votes and election results.

2.3. Public health risks and gender-based violence (Article 34(1)d)

As it became clear during the Covid-19 pandemic, medical disinformation has profound and immediate consequences, not only for individual well-being but also for public health at large. False or misleading health claims can lead to harmful behaviors, discourage evidence-based medical treatments, and contribute to the spread of misinformation that undermines trust in healthcare institutions.

An important aspect of medical disinformation is its frequent link to monetization. For instance, Fundación Maldita.es has repeatedly warned its community about how deepfakes

of legitimate doctors are promoted through Meta Ads in order to redirect to websites where unregulated or ineffective medical products are offered. These fraudulent endorsements exploit the perceived authority of medical professionals to persuade users into purchasing potentially harmful treatments or supplements.

Another concerning trend in health-related misinformation is the promotion of eating disorders, particularly on TikTok, a platform with a predominantly young and thus vulnerable audience. Certain content subtly or overtly encourages disordered eating habits, glamorizing extreme dieting and unhealthy body image standards. Given the vulnerability of this demographic, such content can contribute to serious mental and physical health issues.

In the same platform and collecting thousands of views, Fundación Maldita.es has gathered endless examples of trends, challenges or clips that systematically promote violence against women. These TikTok videos openly promote misogynistic ideas, sparking strong reactions both in support and opposition. This high engagement signals the algorithm to boost their visibility, ultimately increasing their reach.

Platforms like Facebook, Instagram, and TikTok are not only hosting and pushing problematic content that impacts women but also allowing advertisements for harmful applications. Specifically, ads promote AI-powered apps that generate non-consensual sexualized content of celebrities, such as fake videos of them kissing or digitally altered images showing them nude or in revealing clothing. For example, we identified a campaign of 150 ads on Instagram and Facebook promoting one such app. Similarly, TikTok featured a series of ads containing AI-generated videos of celebrities kissing, further amplifying this issue.

2.4. Illegal Content and Fundamental Rights Violations (Article 34(1)a-b)

Fundación Maldita.es has studied two distinct risks directly linked to X's systems that enable the spread of illegal content, particularly violations of the right to one's image, scams, and hate speech.

First off, a campaign of over 165 promoted posts on X was analyzed. Blue-checked accounts exploited the image of well-known Spanish celebrities without consent on X's ads in order to attract users and redirect them to web pages that offered false investments in cryptocurrencies. In less than three months of data collection, the scams took advantage of the ads systems reaching over 368,000 users on average and being viewed by a minimum of 76 million times overall.

Following the launch of Grok's free access, a different risk surfaced. The AI model integrated into X was exploited to generate violent, humiliating, and sexualized images of political and public figures, often with racist or xenophobic connotations. This was made possible by the lack of safeguards within the AI system to prevent such misuse.

3. Mitigation Measures by Online Platforms

In response to harmful misinformation and disinformation, context-adding interventions such as fact-checking labels are particularly effective, as they help users recognize the significance and magnitude of such content while scrolling through the platform. However,

VLOPs and VLOSEs address misinformation and disinformation risks in other ways that involve a range of visible actions on particular content.

As part of the above mentioned Elections24Check project, Fundación Maldita.es built upon the database of posts containing debunked misinformation to assess the visible response of online platforms hosting them. While the response rate was very different comparing all analyzed platforms, Facebook, Instagram, X, TikTok, and YouTube, the following were the actions taken against disinformation and visible to users in most repeated order and with an evaluation of its effectiveness:

3.1. Fact-checking labels

In the mentioned report about action on debunked EU-Elections disinformation, Meta's Facebook (88,83%) and Instagram (70,73%) had the highest response rate, fact-checking labels being their top type of moderation applied.

These tags, result from Meta's program in collaboration with independent fact-checking organizations, provide debunking information directly alongside disinformation claims. This helps users make informed decisions about whether to reshare content while fully respecting freedom of speech.

As outlined in this position paper, scientific evidence shows that fact-checking labels effectively reduce both the spread of misinformation and people's perception of a post's truthfulness, more so than other types of labels. Their effectiveness has been demonstrated across various topics and countries, thanks to fact-checking organizations' rigorous methodologies and local expertise.

3.2. Community Notes

While fact-checking labels count on the evidence collected by independent professional organizations with verified standards and methodology, community notes rely on regular users to provide relevant information and an algorithm to assess whether enough 'consensus' has been reached in order for the label to be displayed.

This crowdsourced mitigation strategy is potentially effective in combination with other initiatives but has been proven not reliable enough by itself to address online disinformation. During the EU Elections, only 15.8% of the posts in X flagged by European fact-checkers as containing debunked disinformation had a visible community note. A similarly low figure was observed during the floods in Spain, when 8.5% of the posts with harmful disinformation had a visible note. In this case, we saw that 1 out of 4 posts with no visible note had a proposed note not being displayed. This is a reflection of a wider challenge: in 2024 only 8.30% of all over 1 million notes suggested by users were visible on X.

More positive results appear if notes citing the work of fact-checking organizations are analyzed. 12.06% of notes with a link to an international fact-checking organization (IFCN signatory) became visible, 15.23% if European fact-checkers (EFCSN verified members) were considered. Similarly, this group of notes are faster in becoming visible (90 minutes earlier than general notes).

These organic results show the opportunities of combining systems such as community notes with the work of independent fact-checking organizations. All the ability of users to detect dangerous disinformation and the ability of fact-checkers to verify that the sources are of quality, add more when necessary and try to get the notes visible as quickly as possible.

3.3. Removal

Removal was the third most common visibility restriction in the analysis of the EU Elections, and for disinformation posts on TikTok, it was the most frequently used action.

TikTok works with fact-checking partners to assess flagged videos potentially containing misinformation. However, unlike Meta, which integrates fact-checking information directly into posts for users to see, TikTok primarily removes misleading content. This severe action was nonetheless never taken on many videos flagged by Fundación Maldita.es and collected millions of views during the 2024 floods in Spain.

Even if YouTube's policy on disinformation also foresees removal as the main moderation decision, only 4% of the videos debunked by European fact-checkers during the elections had been taken down by the platform reflecting inconsistency in their moderation efforts.

While removing illegal content is necessary, this approach is less effective for addressing misinformation. Deleting lawful but disinformation content without providing context can foster distrust among users hindering freedom of speech. It also eliminates the opportunity for users to engage with fact-checked information and make informed decisions themselves, an empowerment that labeling provides but removal does not.

3.4. Generic information panels or labels

Beyond Meta's fact-checking labels, platforms like YouTube and TikTok also use various panels and labels to improve access to reliable information.

YouTube, for example, adds information panels to videos covering topics prone to misinformation, such as climate change. Rather than addressing specific claims, these panels provide basic information from authoritative sources like the United Nations. Additionally, YouTube marks media outlets under video descriptions if they are state-funded or government-owned.

TikTok, on the other hand, creates election information centers that users can access through labels on relevant content. These centers, developed in collaboration with local organizations, offer verified details about the electoral process and results. In times of crisis, TikTok sometimes displays search prompts like, "Learn about natural disasters and response plans from reliable sources."

X occasionally applies generic labels to manipulated or out-of-context content, but instead of providing further explanation, these labels simply redirect users to the platform's policies.. More recently, several platforms have introduced labels to disclose synthetic content, either through automatic detection or user reporting.

Despite these efforts, significant gaps remain. Fundación Maldita.es has analyzed YouTube's failure to address a network of politically misleading channels using synthetic

audio. These channels, with over 400,000 subscribers, continue to spread disinformation unchecked.

Overall, labels are most effective when they are specific and directly relevant to the content they accompany. When designed to genuinely improve access to authoritative information, they can serve as a valuable complement to other initiatives combating misinformation.

4. Risk Factors

4.1. Blue checks on X

Reaffirming the European Commission's basis for opening formal proceedings against X, one of the platform's most significant risk factors is its blue check system.

Originally designed as a verification tool for notable users, the blue check now simply indicates a paid subscription. However, many users still associate the symbol with trustworthiness, assuming it confirms the account's authenticity. This shift is highly deceptive, as it undermines the original purpose of verification, making it easier for bad actors to appear credible.

Additionally, blue check subscribers receive visibility boosts, making the system particularly attractive to those seeking to amplify their reach, including those spreading harmful content. Fundación Maldita.es investigated scams using celebrity images on X and found that every account in the database carried a blue check. Similarly, during the floods in Spain, nearly half of the accounts sharing debunked disinformation were verified subscribers, highlighting how the system is being exploited to spread disinformation claims more effectively.

4.2. Advertisement systems and automated moderation

Regarding the recurring systemic risks identified, advertising systems on platforms like X, Meta, and TikTok have enabled the widespread circulation of harmful content, often reaching thousands or even millions of users. Problematic ads, such as those promoting scam websites, deepfakes of doctors, or AI-generated sexualized videos, frequently bypass automated content review systems. These systems, designed to screen ads before publication, are consistently exploited by bad actors.

Targeted advertising further amplifies these risks, allowing ads to be directed based on age, gender, or other demographic criteria, particularly concerning when promoting medical products.

Fundación Maldita.es has documented the use of cloaking techniques to evade detection in such cases. Additionally, Meta Ads' dynamic product services make it possible to slip problematic content through by embedding deepfakes among seemingly non problematic product images. According to Meta, their system "can automatically create the right combination for the audience," meaning different users may see different versions of the same ad, some of which may violate community standards.

Content moderation efforts on these platforms seem focused on reacting rather than preventing this content, often only addressing harmful ads, if ever, after they have already garnered significant reach and impressions.

5. Other information

The risks posed by very large online platforms are often interconnected and not limited to those not reporting over the users' threshold. Smaller platforms also play a significant role in the spread of harmful content. For example, Fundación Maldita.es, in collaboration with the University of Granada, has analyzed how public Telegram channels contribute to the dissemination of disinformation.

Compared to traditional media channels, so-called "alternative" channels on Telegram often have a wider reach and generate higher engagement. They build strong and active communities around harmful content, reinforcing narratives through continuous interaction on comments. Features like the 'Similar Channels' tab further amplify this issue, creating a loop that pushes users deeper into networks of disinformation.

These channels do not operate in isolation in Telegram but rather interact with other platforms to expand their influence. They attract followers from external platforms while simultaneously directing their audiences to content hosted elsewhere. For instance, our analysis found that over 80,000 messages in these 97 channels contained links to YouTube, illustrating how disinformation spreads across multiple digital ecosystems.

6. Conclusion

The European Board for Digital Services and the European Commission should strive to produce a report under Article 35(2) that is both comprehensive and appropriately detailed, taking into account the broad scope of services it addresses and the unique contexts of the contributing Member States.

Regarding best practices for mitigating disinformation, Fundación Maldita.es recommends that the Board aligns its approach in the upcoming report with its opinion on the recently converted Code of Conduct on Disinformation. Collaboration with independent fact-checking organizations is one of the pillars of the Code, which now becomes a benchmark for effective enforcement under the Digital Services Act (DSA).