Why Professional Fact-Checking Matters: Meta, the EU, and the Western Balkans

ZaštoNe

OPEN SOCIETY
FOUNDATIONS
WESTERN BALKANS

The creation of this text was made possible with the support of the
Open Society Foundations – Western Balkans within the project Growth and Digital Rights:
Ensuring the Integration of the Western Balkans into the Digital Single Market.
The content of the text is the responsibility of the authors
and does not necessarily reflect the views of the
Open Society Foundations – Western Balkans.

PARTNERS
SERBIA

Institute
for Democracy
and Mediation

WOMEN
in tech
KOSOVO

NVO 35 mm

On 7 January 2025, Meta's founder and CEO Mark Zuckerberg announced the company's decision to end its Third-Party Fact-Checking program (TPFC) on Facebook, Instagram, and Threads in the United States. The move was part of a broader overhaul of Meta's content moderation policies, presented as an effort to "prioritize free speech" and "combat censorship". Starting in the U.S., Meta decided to replace cooperation with fact-checkers with a community-based labelling system called Community Notes, similar to the one already in use on Twitter, now renamed to X.

Launched in 2016, the TPFC is based on cooperation with independent fact-checkers certified by the International Fact-Checking Network (IFCN) and, as of 2024, by the European Fact-Checking Standards Network (EFCSN). Fact-checkers review and rate the accuracy of content through original reporting, published on their respective websites. Meta's subsequent actions were designed to ensure that users were informed about disinformation and that fewer people saw such content by applying labels and, in some cases, restrictions, but without removing the content itself. When content is rated by fact-checkers, the platform adds a warning label linking to the fact-checker's article, so users can read additional information and context. Content rated as false also receives reduced distribution and is not suggested in user feeds.

Community Notes, on the other hand, relies on contributions from eligible users who submit notes on posts they consider misleading or lacking in background information or context. As Meta explains, once a contributor submits their community note, other contributors can rate whether a note is helpful. A note gets published if there is agreement that it is helpful "among contributors that normally disagree with each other", and this is established by taking into account "each contributor's rating history". In addition, community notes carry no penalties: they do not reduce distribution, affect visibility, or restrict sharing.

For now, the new system is being tested only in the U.S. To this date (September 2025), Meta has not announced plans for the rest of the world, though it had previously suggested it would evaluate and improve the approach over the course of the year before expanding to other countries. Still, a global end to the program would not come as a surprise, given the highly politicized context in which the decision was made.

What about Europe?

The impact of these policy changes in Europe remains uncertain. In his January statement, Zuckerberg criticized Europe for having "an ever-increasing number of laws institutionalising censorship", clearly alluding to its digital services regulations. Shortly after, Meta's chief lobbyist Joel Kaplan said that the program will launch "elsewhere" in 2026, potentially including the EU, but stressed that the company would work with European regulators when making such changes.

Under the EU's Digital Services Act (DSA), the largest online platforms - including Facebook and Instagram - can be held legally accountable for the spread of disinformation. However, the DSA is not primarily concerned with disinformation and does not require platforms to act on individual instances of disinformation, or any other specific piece of content. Rather, it relies on preventive and reactive approaches and efforts to combat and mitigate the risks and harms of "lawful but awful" categories of speech, such as disinformation. The DSA's risk-based approach relies on heightened due diligence obligations for very large online platforms and search engines (VLOPs and VLOSEs). These include the obligation to "identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems", which may also relate to the design of their systems for content moderation and recommendations, and selection and presentation of advertisements (Article 34 DSA). VLOPs and VLOSEs are thus obliged to assess how their business models and design features contribute to systemic risks such as "actual or foreseeable negative impacts" on civic discourse, public health, electoral processes and public security. Importantly, assessments must also consider whether these risks could be amplified by design features that allow for inauthentic use and rapid dissemination of content that is either illegal or incompatible with platform's terms and conditions.

The relevance of the above elements for disinformation, and the way it is being disseminated and amplified on social media, is self-evident. Algorithmic amplification of disinformation clearly represents a systemic risk, as does the inauthentic use of service, such as the use of bots or fake accounts.

When systemic risks are identified, VLOPs and VLOSEs are obliged to put in place "reasonable, proportionate and effective mitigation measures". This requirement is mandatory, but the list of measures proposed by the DSA is quite broad and non-exhaustive: from adjusting their terms of service, content moderation processes and algorithmic systems, to adopting targeted measures aimed at limiting the presentation of advertisements (Article 35 DSA).

So, how does fact-checking fit in here?

One of the possible risk mitigation measures includes cooperation with other online platforms through the codes of conduct, whereby platforms voluntarily agree on commitments to take specific risk mitigation measures. The EU launched its first Code of Practice on Disinformation in 2018, further strengthening commitments in 2022. Initially, the Code was a purely self-regulatory mechanism and it attracted a broad range of major tech companies that agreed to implement various measures, including the cooperation with independent fact-checkers.

On 1 July 2025, the Code of Practice was formally integrated into the DSA framework as a Code of Conduct on Disinformation. While still voluntary, it has now been elevated into an official co-regulatory instrument under the DSA, meaning that commitment and compliance with the Code count toward addressing systemic risks linked to disinformation. Theoretically, if online platforms that have signed the Code fail to comply with its provisions, they could face enforcement actions under the DSA, as their commitments will be subject to scrutiny through the mandatory independent audits required by the DSA.

Empowering the fact-checking community has been one of the six key areas since the Code was first launched in 2018, and it remains so in the updated Code of Conduct. Signatories committed not only to cooperating with independent fact-checkers and using their work in their services, but also to extending fact-checking coverage across all EU Member States and languages, and to providing fair financial contributions to European fact-checking organizations for their work.

However, once the Code stopped being purely voluntary, platforms began retreating from their commitments, especially those related to fact-checking. Whereas X left the Code altogether following Elon Musk's acquisition in 2023, Google (including Search and YouTube) and Microsoft (Bing and LinkedIn) withdrew from the fact-checking chapter in early 2025. TikTok, which launched its Global Fact-Checking Program in 2020 and has been expanding it since, has made its commitments conditional, stating it would uphold them only if other signatories offering similar services did the same. In the event that Meta withdraws, it is likely that TikTok will follow suit. While maintaining its collaboration with fact-checking organizations, TikTok has already started testing a new feature called "Footnotes" in the U.S., which allows users to add context to videos through community-sourced notes.

Still, even platforms that did not sign the Code are required under the DSA to undertake measures against disinformation. Codes of conduct serve as benchmarks for the European Commission to assess whether platforms are fulfilling their due diligence obligations. Platforms are not obliged to participate in a Code or subscribe to all of its commitments, but if they choose not to, they must demonstrate that the measures they adopt instead are as effective at addressing systemic risks.

In other words - the regulatory mechanism under the DSA does not mandate cooperation with independent fact-checkers. Platforms remain free to decide which measures they will put in place, as long as they manage to effectively mitigate the risk. The question is: without the input of professional fact-checkers, can disinformation-related risks truly be mitigated?

Professional vs. crowdsourced fact-checking

In its open letter to Mark Zuckerberg, the IFCN warned that Meta's decision to end the program worldwide would be "almost certain to result in real-world harm in many places", given that the program extends to many countries "highly vulnerable to misinformation that spurs political instability, election interference, mob violence and even genocide". Concerns have also been raised about state-sponsored disinformation campaigns, arguing that user-based content verification will be unable to identify coordinated campaigns and tactics used to spread false content. Another issue relates to the monetization of disinformation: under Facebook's current content monetization policy, content rated false by a third-party fact-checker is ineligible to monetize. As ProPublica noted, this rule will become irrelevant once Meta stops working with fact-checkers. This is especially concerning given Facebook's new content monetization program, which significantly broadens the range of content eligible for monetization and could potentially create more financial incentives for amplification of hoaxes.

In its statement on Meta's announcement, AlgorithmWatch noted that cutting fact-checkers and replacing expert analysis with "armchair expertise" will make reliable information harder to access and will "almost certainly" facilitate disinformation. Researchers and experts have repeatedly questioned the effectiveness of crowdsourced fact-checking compared with professional fact-checking, emphasizing that while it could be a useful complementary tool, it cannot serve as a substitute.

On X, Community Notes are only displayed when they reach consensus among users with different (political) views. A proposed note becomes visible if it is rated as "useful" by people who have previously voted differently on other notes. Meta's approach is similar, as its system also relies on a contributor's rating history. According to Meta, "this approach helps ensure that notes reflect a range of perspectives and reduces the risk of bias".

However, research suggests that this design prioritizes cross-partisan agreement over factual accuracy. As a result, disinformation on contentious issues often goes unflagged simply because users cannot agree - even when the content is inaccurate. Community Notes may also remain hidden or take a long time to appear before consensus is calculated. When it comes to highly polarizing topics, where consensus is not likely to be reached, inaccurate content may remain unflagged. An analysis done on 1.7 million Community Notes on X found that fewer than 10% of submitted Community Notes are published. Research by the Poynter Institute during the 2024 U.S. Presidential Elections reached a similar conclusion, finding that Community Notes had only a marginal impact on election information quality.

In the EU, an investigation by Science Feedback showed that only 51,640 Community Notes (approximately one in 500,000 tweets) were displayed in the first 10 months of 2024, a period including the European Parliament elections. Cross-referencing with the database of content reviewed by the EFCSN's members around the EU elections revealed that as much as 68.8% of content on X that fact-checkers found to be false or misleading received no intervention. Using the same database, the Spanish fact-checking organization Fundación Maldita observed that X took no visible action in 70% of cases, and that Community Notes appeared in only 15% of posts already debunked by EFCNS members during the elections.

Fact-checking labels play an important role in empowering users to make informed decisions about the content they view, interact with and share. By providing valuable context, fact-checking reduces the credibility of disinformation, making users less likely to share such content. In its statement following Meta's announcement, the EFCSN pointed out that Meta itself had previously praised its Third-Party Fact-Checking program and the effectiveness of its labeling system, as well as the benefits for its users. Ahead of EU Elections, it stated the following: "Between July and December 2023, for example, over 68 million pieces of content viewed in the EU on Facebook and Instagram had fact checking labels. When a fact-checked label is placed on a post, 95% of people don't click through to view it."

Indeed, in another study analyzing the impact of fact-checking in X's Community Notes, Fundación Maldita found that notes citing evidence from a fact-checking organization are more trusted by users and therefore more likely to become visible. They also appear alongside the original tweet much sooner, 90 minutes earlier than general notes. This speed can be crucial, as research suggests that Community Notes are often too slow to curb engagement with disinformation during the early and most viral stage of its dissemination.

The model that has been established by Meta's TPFC program has also produced some tangible results that can not be replicated through Community Notes and are especially significant for regions like Western Balkans, where media markets struggle with both sustainability and credibility.

The impact of TPFC in the Western Balkans region

Before the TPFC expanded to the region, the incentives created by VLOPs have had an overwhelmingly negative effect on information integrity in WB, pushing the media towards clickbait, sensationalism and other information manipulation as tactics to reach online audiences. The restrictions introduced by the TPFC - demotion of fact-checked content labelled as false or misleading or, in case of "repeat offenders", demotion of entire pages on Meta's platforms - have influenced a change in behavior of many outlets that rely on online revenue. In order to avoid the "penalties" set up by Meta's TPFC, some moved towards higher standards of verifying information before publishing and others towards issuing timely and comprehensive corrections, since publishers whose content has been labelled as false can restore the status of their posts and pages if they correct it in line with Meta's standards.

This principle alone has fostered a practice of correcting false information in dozens of media outlets that rarely or never issued corrections of fact-checked claims before the program was introduced. While there is no comprehensive data on the number of corrections resulting from the TPFC program, individual cases offer valuable insights into the scope of this impact.

Sarajevo-based Raskrinkavanje, for example, entered the TPFC program in September 2020 and has since documented corrections of 4433 media articles and social media posts they have fact-checked and rated on Meta's platforms. Additional 413 media articles fact-checked by Raskrinkavanje before September 2020 were retroactively corrected by some publishers. Together, that amounts to almost 5000 pieces of false or misleading claims that have been transparently and thoroughly corrected as a result of just one fact-checker's participation in the TPFC program. Similar impact can be traced in other countries in the region as well. If the TPFC ends in the region, this practice will likely end with it.

What is next?

The EU's commitment to supporting and maintaining professional fact-checking is, at least nominally, strong. Even though DSA enforcement has not yet provided concrete answers, it does offer legal avenues for holding platforms accountable in applying effective systems for fighting disinformation. For example, in December 2023, the European Commission initiated formal proceedings against X under the DSA regarding, among other things, the effectiveness of the Community Notes system and related policies mitigating risks to civic discourse and electoral processes. In the EU, it can therefore be expected that the effectiveness of measures taken by online platforms to combat information manipulation will be subject to regulatory scrutiny.

The European Democracy Shield is the latest initiative of the European Commission dedicated to strengthening information integrity in Europe. It focuses on four key areas: countering disinformation, strengthening information integrity in election campaigns, enhancing societal resilience, and boosting citizen engagement. Expected to be adopted and published in November 2025, this initiative recognizes that fact-checkers play a crucial role in countering disinformation and in providing valuable insights for DSA risk assessments. Furthermore, the EU has expressed its commitment to ensuring long-term conditions under which they operate. In its call for proposals to strengthen the European network of fact-checkers - launched in line with the European Democracy Shield's objectives - the Commission explicitly included stakeholders from candidate and accession countries, as regions that are particularly vulnerable to disinformation and foreign interference. If this is understood as a signal that the protections envisaged under the European Democracy Shield will extend to Western Balkans as candidate countries, that would be a welcome development.

For candidate countries, aligning early with the EU's regulatory and policy framework is therefore not only a matter of preparing for accession, it is the most realistic path to ensuring resilience against disinformation and protecting the integrity of democratic processes.

Currently, the moderation of content on very large online platforms in the region is regulated solely by the platforms' internal rules, defined in their terms of service and community standards. These platforms are not yet required to comply with the provisions of the DSA, which means there is no incentive to implement the Code of Conduct on Disinformation in the Western Balkans. Facilitating alignment with EU digital policy is therefore the best tool currently available, and regional governments should step up their efforts in this regard. Information integrity should be positioned higher on the list of priorities in the EU accession process.

The support of the European Union will be crucial in this regard, particularly in ensuring cooperation between very large online platforms and local institutions, and in enforcing measures to mitigate systemic risks such as those related to disinformation. Such cooperation is possible even in the absence of formal regulations. Moldova, another EU candidate country, offers a good example: recently, the EU and Moldova deepened their digital cooperation to strengthen resilience in strategic areas. Among other things, with the Commission's support, a new EDMO hub was established in Moldova.

In the event Meta decides to end cooperation with its fact-checking partners from the region, the risks of disinformation becoming further entrenched in the region will grow. The shortcomings of consensus-based models, such as Community Notes, underline the importance of professional fact-checking as an accountability mechanism.

The European Union's next steps will therefore be closely watched across the Western Balkans, where its support and backing could prove decisive in countering disinformation and strengthening democratic resilience.

Maida Ćulahović and Tijana Cvjetićanin