Putting X's Community Notes to the Test

VB verfassungsblog.de/putting-xs-community-notes-to-the-test/

Marc Bovermann

08 January 2024

All of the biggest social media platforms have a problem with disinformation. In particular, a flood of false information was found on X, formerly Twitter, following the terrorist attack by Hamas on 7 October 2023 and the start of the war in Ukraine. The EU Commission therefore recently initiated formal proceedings against X under Art. 66 para. 1 of the Digital Services Act (DSA). One of the subjects of the investigation is whether the platform is taking sufficient action against disinformation. If it fails to do so, it is in breach of its risk mitigation obligation under Artt. 35 para. 1, 34 para. 1 subpara. 2 lit. c DSA. Consequently, the Commission would require the platform to submit an action plan for risk mitigation, according to Art. 75 para. 2 DSA. Pursuant to Art. 74 para. 1 lit. a DSA, the platform also faces a fine.

Despite these stakes, X takes an approach different to all other platforms: As can be inferred from the X Transparency Report dated 03.11.2023 posted information is not subject to content moderation, but solely regulated through a new tool: The Community Notes. In particular, this means that X does not check the content of its services for false information through algorithms or trained personnel. This approach does not only save costs, but is also fully in line with X's new understanding of its role in the public discourse, following Musk's ("free speech absolutist") takeover. The Community Notes seem to be an integral part of this new understanding. But in the end, they are not enough to stop disinformation.

Community Notes as a Panacea for Disinformation

Community Notes are short assessments by other users about potentially misleading or incorrect tweets. These assessments are displayed as disclaimers next to such tweets, in order to inform users about missing context. Notes are drafted and discussed in a backend platform, separate from the main platform. Here, contributors can exchange information anonymously and draft Notes. Helpful Notes can be rated positively, unhelpful Notes negatively. The answer "partially helpful" is also possible. If enough contributors rated a Note helpful, it will also be displayed on the main platform. However, the decision to select a Note can be overturned later.

The tool also differs significantly from the main platform in regard to other design features: The source code of the service is available to the public on GitHub in order to increase transparency. Furthermore, whether a Note should be displayed publicly is not determined based on majority rules, but on a so-called bridging-based ranking system. Contributors are

assigned a certain perspective based on their previous voting behaviour. A Note is only displayed on the platform if it receives enough positive votes from people with different perspectives. This means, people who disagreed in the past must agree on a specific Note. Only Notes that are neutral in tone and correct in content should thus be displayed on the main platform.

Lastly, X made some changes as to have Notes appear on the main platform more quickly. This ensures that more users see the clarifying Notes. The platform also stopped monetisation of tweets with a Note. This prevents users from earning money with sensational or misleading statements.

The feature was introduced in January 2022, still under the name of its predecessor "Birdwatch", with which Twitter had some success in combating disinformation. Following Musk's takeover in November 2022, the feature was quickly scaled up. By the end of November 2023, the option to report misleading tweets under X's Terms and Conditions was fully removed. Thus, the Community Notes have no choice but to take on the role of a panacea in X's fight against disinformation.

To say that the tool is a panacea would probably be an understatement for Musk. He believes that his platform provides the most accurate information on the entire internet. Consequently, not only the Community Notes themselves are put to the test but also the new understanding of the platform itself.

Platform's Duty to Combat Disinformation

An important preliminary question for the Commission is whether X has an obligation to take action against disinformation at all and, if so, what such action might look like.

Due to Art. 33 para. 1 DSA, a duty to combat disinformation only applies to very large online platforms (VLOPs). X also belongs to this select group. However, it is worth taking a closer look at the regulations to find out how the DSA formulates this obligation in detail.

This obligation cannot be inferred from the text of the regulation itself. Instead, reference is made to Art. 35 para. 1, 34 para. 1 subpara. 2 lit. c DSA, which is an obligation to mitigate systemic risks to civic discourse. A glance at the recitals quickly confirms that disinformation constitutes such a systemic risk (recitals 83, 84, 88, 95, 104, 106). However, only recital 104, which actually serves to explain the concept of co-regulation, explicitly states that disinformation also poses a systemic risk to democracy. If Art. 34 and 35 DSA are interpreted in the light of this recital, a duty to combat disinformation is constituted – no problem at all, one might think.

However, this does not answer the question of what exactly the risk mitigation should look like. The DSA does not take an explicit position on this. However, according to Recital No. 2, combating online disinformation is a key objective of the DSA. Disinformation is also a prime

example of the social risks posed by VLOPs. Against this background, it therefore seems surprising that the DSA does not take a clear stance on the issue, not even within the recitals. Recital 82, which is intended to explain the systemic risks for civic discourse in more detail, the word ‘disinformation’ does not appear.

The Commission’s answer refers to the “Code of Practice on Disinformation” (CoP) instead. The CoP was adopted by the Commission in 2022, still with the involvement of Twitter. It contains many obligations, some of them very specific, that the participating companies have imposed on themselves in order to effectively combat disinformation on their platform.

Codes of conduct (such as this one) are voluntary according to the concept of the DSA in Art. 45 para. 1. However, this is only true in theory for VLOPs, as already indicated by recital 104: Whether a VLOP has taken sufficient risk mitigation measures is, among other aspects, determined by whether it participates in a code of conduct. Otherwise, it is difficult to prove that it has fulfilled its obligation to mitigate systemic risk. The legislator has even codified this interplay: The Commission takes a voluntary commitment into account when drafting action plans in accordance with Art. 75 para. 2 DSA. Ultimately, it also appears difficult for a VLOP to sufficiently mitigate its systemic risk without simultaneously adhering to a corresponding code of conduct.

Therefore, if VLOPs identify a systemic risk for which there is a code of conduct, compliance with it is de facto mandatory. Thus, the CoP therefore also applies to X – whether voluntarily or involuntarily. For Musk, who previously ended his participation in this very code of conduct, this means: Welcome back!

The Issues of the Community Notes

The Community Notes are intended to fulfil the obligation to mitigate systemic risk to civic discourse. But are they indeed the panacea for disinformation Musk has promised?

Little can be said about the half-life of a tweet other than that it is very short. This means that most users see a tweet only in a short time after it is published. A discussion and vote on Notes trying to correct information in a tweet cannot possibly take place during this time. The majority of users therefore pick up on incorrect information without ever having seen a potentially clarifying Community Note. This is a basic structural problem caused by the ‘accessory’ status of Community Notes.

To compensate for this, X has set up a system that notifies users who have interacted with a tweet that was subsequently provided with a Community Note. Only users who have interacted with a tweet (e. g. like, retweet, comment) are notified. Users who have only read the tweet are not notified. It is worth noting that the number of users who have only read a tweet but not interacted with it is usually much higher than the number of interacting users.

This restriction to interacting users seems questionable. Art. 33 para. 1 DSA, concerned with awarding the status “VLOP”, provides an answer. The status is awarded to an online platform based on a very high number of “active users”. With regard to the question of which users are “active”, Recital 77 explicitly opposes to merely include users who interact with information on the platform. The same goes for the CoP, which only focusses on the “visibility” of disinformation. Users who only view the content on the platform are therefore also “active”. In the end, they are also exposed to the social risks of VLOPs. A mitigation of the risk for the aforementioned user group is therefore not reliably achieved.

The bridging-based ranking is also a welcome feature. However, it is worth noting that the system is actually intended to be a counter-model to engagement-based ranking. The latter is based on a core principle of platform economics that underlies every recommendation system of VLOPs, including that of X: Its goal is the greatest possible interaction with a tweet and maximisation of time spent on the platform. The fact that a new ranking system is now being introduced for the Community Notes, which is intended to compensate for the deficits of the ranking system actually used, is like trying to treat a broken leg with a plaster.

Furthermore, the Community Notes are also exposed to manipulation by users. The platform does little to prevent users from manipulating the ranking system by coordinating their voting behaviour. These users often have more than one account. The platform could counteract this by applying stricter criteria when selecting users and by moderating the Community Notes. The same also applies to existing accounts; X could also apply stricter standards here.

Users are also responsible for selecting the posts that are to be provided with a Community Note. A focus on political content was identified, although political content only accounts for a small proportion of the information on the platform. Certain accounts (such as Elon Musk’s) are also particularly affected by Community Notes. Combined with the possibility of targeted manipulation of the service through collusion between contributors, the system risks being at the mercy of the political interests of third parties. The platform could counteract this by curating the selection of Notes that a contributor can rate in such a way that the risk of manipulation or harassment of a particular person is mitigated.

Ultimately, it must be criticised that the platform has stopped moderating false information. The Community Notes are a welcome measure to combat disinformation. However, they cannot replace a functioning moderation system. The CoP also conceptualises the involvement of users only as part of a larger overall concept. Without a moderation team within X Corp., the goal to combat disinformation has no place in the corporate structure. X should bear in mind that systemic risks also require systemic responses.

Putting the DSA to the Test

This is the Commission's first formal procedure under the DSA. It is not only the Community Notes that are being scrutinised, but also the Regulation itself. The proceedings address core concerns of the DSA. Bearing in mind the still necessary clarification of the DSA, one must hope for a non-compliance decision under Art. 73 para. 1 (in conjunction with Art. 75 para. 2) DSA, as to have the Commission outline the VLOP's obligations under the DSA.

This is especially true for the unclear systemic risk assessment and mitigation obligations. One foundational dilemma that emerged when the CoP was concluded is particularly interesting: VLOPs are obliged to combat legal content that contains false information under Artt. 35, 34 DSA. At the same time, they are also bound by fundamental rights under Art. 14 para. 4 DSA; an expression of opinion can therefore not simply be removed from the platform. This conflict has so far been resolved in favour of "free speech" under Musk's control. This approach is generally deemed a non-starter in the EU. But it points to deeper conflict emerging among the different notions of freedom of speech around the globe. Whichever way the Commission decides this conflict in substance, it is where it can prove that the DSA is tenacious and clearly show that the digital space is not simply left to Musk's notion of civic discourse.

A Conclusion in 280 Characters

The Community Notes are no panacea. X has therefore not sufficiently fulfilled its obligation to mitigate systemic risk. It is now up to the Commission to reconquer the (digital) public sphere on X.

SUGGESTED CITATION Bovermann, Marc: *Putting X's Community Notes to the Test, VerfBlog,* 2024/1/08, https://verfassungsblog.de/putting-xs-community-notes-to-the-test/, DOI: 10.59704/9449c0ca83d334d3 .

Explore posts related to this:

Other posts about this region:
Europa