

Months before the French election, Facebook gives a free pass to far-right hate

In an increasingly heated political climate, Facebook fails to enforce its own content moderation policies and remove hate posts - including incitement to violence against political candidates, women and migrants - even after respective content that violates platform's terms of service and respective French law has been notified to the platform by users.

TRIGGER WARNING!

**THIS REPORT INCLUDES CONTENT THAT FEATURES
HATE SPEECH AND STRONG ABUSIVE LANGUAGE**

Key findings

- * **In 70 % of the cases, Facebook failed to delete hate comments even after we notified them to the company through their flagging system.** This included insults against women and political candidates (e.g. "*Je chié dans ta gueule espèce de salope*") as well as racist hate speech (e.g. "*race de bâtards, a passer au lance flamme*").
- * **94%** of notified comments that Facebook failed to delete (out of 205 comments), **were assessed by legal experts as violating French law.**
- * **Facebook also failed to handle user notifications diligently and transparently,** indicating profound deficits in Facebook's notice and action procedures. Facebook replied within the 24 hours' time frame in less than 20 % of the cases.
- * **The hate comments had been online from 19 to 690 days** (431 days on average) when we reported them, despite violating Facebook's community standards or French law.

- * The findings suggest that the **threat of credible financial sanctions is needed for Facebook to comply** with existing rules and protect the rights and safety of its users.
- * These findings come at a time when **French democratic representatives receive death threats, raising important questions about the role and responsibility of parties** in encouraging healthy public debates, on and offline.
- * By failing to enforce its own terms of service consistently, Facebook rewards the use of inciting content for political mobilisation and distorts political competition at the expense of those actors who “play by the rules”.
- * These findings also draw attention to the **need for platforms to take more systemic approaches to regulating both manifestly illegal and toxic content online** and carefully consider

the issue of the mainstreaming of hateful content and the tangible implications it can have in the context of elections.

- * This study suggests social media companies must not only double down their efforts to comply with their own moderation policies, **they should also take a cross-harm risk mitigation perspective when developing their products**, so as not to enable such a toxic climate.
- * These findings come **just months before the EU is set to close the negotiations on the Digital Services Act** that will lay content moderation rules for Facebook and platforms alike; they open a serious question about Facebook’s readiness to comply with the forthcoming rules and highlight a **need for a strong enforcement regime**.

Data collection

From the dataset of 2 412 114 public Facebook comments collected by researchers, we selected 280 highly toxic comments drawing on the Perspective API¹. The majority of comments were found below posts associated with the far-right groupings in France (see Figure 1). We assess all of them to be in breach of Facebook’s own community standards², or illegal under the French law.

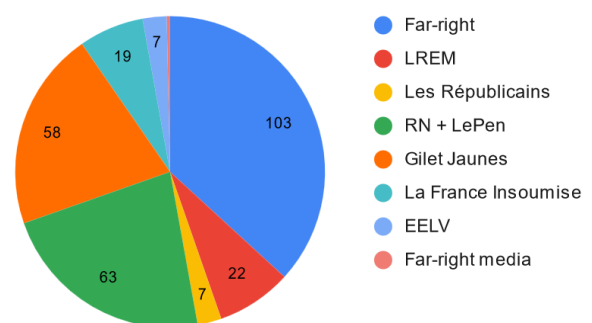


FIGURE 1: DISTRIBUTION OF COMMENTS IN RELATION TO FAR-RIGHT GROUPS

¹ <https://www.perspectiveapi.com/>

² <https://transparency.fb.com/policies/community-standards/>

Monitoring results

Selected 280 comments were reported through the official Facebook reporting mechanism. 88 (31.43%) of the 280 comments reported were deleted after the first day of reporting. After one day Facebook deleted two more comments. However, on the fourth day of monitoring five of the deleted comments were restored. On the fifth day, Facebook restored one more comment. Since the fifth day, there were no more changes. 193 (94%) of the 205 comments that were **not deleted** by Facebook, have been assessed as **violating French law by legal experts**.

After a week of monitoring, only 84 (30.0%) highly hateful comments were deleted

To summarise: after a week of monitoring, only 84 comments (30.0%) containing highly toxic hate speech were deleted.³ These 84 deleted comments had already been online for more than a year (approximately 450 days).

Notice and action procedure

Although Facebook claims that the company will update the notifier within 24 hours of receiving the notification, they **failed to reply on time in 81.78% of the cases**. Facebook had not even created most of the "tickets": we received 60 tickets for 280 reported comments, and only 51 of them were replied to.⁴ Meanwhile, Facebook removed 84 comments, meaning that Facebook failed to inform of their decision to remove a reported comment in 33 cases.

We received three types of replies from Facebook:

1. Facebook agreed to delete the comment, referring to the Community Standards (36 replies), as illustrated in figure 2:

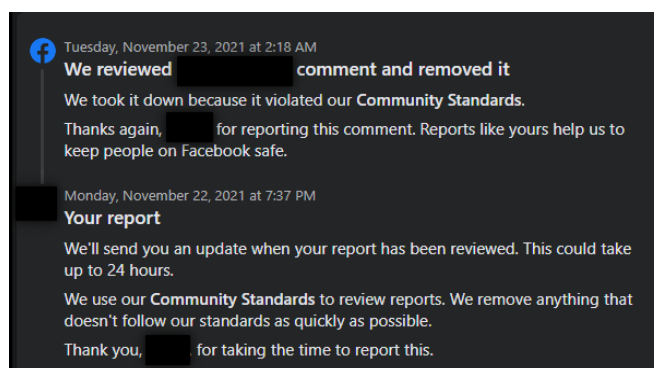


FIGURE 2: RESPONSE TO THE REPORT

³ In the [German Report](#), 50% of the reported comments were removed in 24 hours. This percentage just slightly fluctuated during the whole monitoring period (one week).

⁴ In the [German Report](#), 20.43% of reports have been left without reply.

- Facebook did not agree to delete the comment, referring to the Community Standards (12 replies), as illustrated in figure 3:

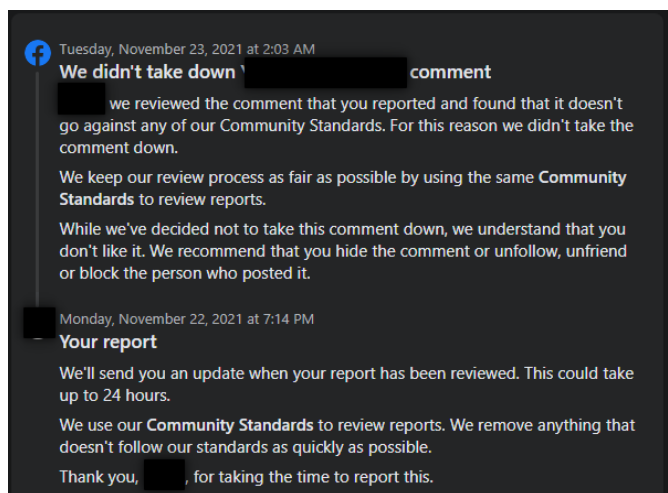


FIGURE 3: RESPONSE TO THE REPORT

- Facebook did not agree to delete the comment, referring to the Technology (3 replies), as illustrated in figure 4:

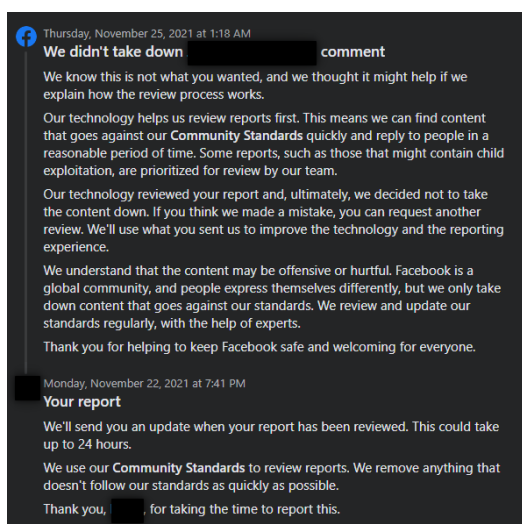


FIGURE 4: RESPONSE TO THE REPORT

In this specific case it is not clear whether any human oversight was involved in making the decision, or the decision was solely made based on the “technology” that is essentially an artificial intelligence. It should draw further attention on quality and human oversight in the content moderation to prevent negative effects on freedoms and rights of users. Human oversight in all steps of the automated process is essential to provide a safety net for the rights of affected users⁵.

⁵ Llansó, van Hoboken, Leerssen, Harambam, “Artificial Intelligence, Content Moderation, and Freedom of Expression,” Transatlantic Working Group, 2020: https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf

Examples of comments that Facebook deleted after reporting

“Petite merde sans couilles qui s en prend a une femme comme d hab retourne ds ton pays le bicot ou la tu te feras violer par tes oncles cousins et toute ta famille d illetres fdp de que tu es et les gens sont tous aveugles putain un truc comme ça je lui defonce la gueule batard de gnoule”

“Il traîne dehors sans masque aussi qu’il chope le Coronavirus et qu’il crève gros bâtard fils de pûte Lallement l’enculé de sa race.!”

“Déchet de père en fils, on devrait les abattre tout les deux ! non, d’abord les laisser pourrir dans une cellule sans rien à manger ni à boire puis après les exécuter ! “

“Le virus c est vous on va vous écraser très bientôt on va redescendre dans la rue et là ça va saigner bande de merde”




“Negre de Maison!!! Honte à toi Negre de Merde Negrion.”

“une balle dans la tête à cet enfant de putain islamiste”

“Steven Bourg sale anti-flic de merde que vous êtes je vous souhaite de crever du corona virus”

“La macronie n à plus lieu d être ni d ailleurs tous les éculés autour aux armes citoyens il va falloir aller les chercher ces filles et fils de putes et les pendre haut et court...”

Examples of comments that Facebook did not delete after reporting

Butez ces merdes de black block   2 Sale pute on va te faire les fesse nous salope  2

Like · Reply · See Translation · 30w

Like · Reply · See Translation · 1y

Sale fils de pute de macronnard. Va niquer ta mère P'tite merde de boudjellal.

Like · Reply · See Translation · 1y · Edited

Et une balle dans ta tronche de pédé , ce serait suffisant ??????

Faut les foutre dans le feu ces primates, sales races


1 r.



Like · Reply · See Translation · 1y


Aller grosse pute, va te raser le cul et met toi la au cul ta médaille connaissance

Like · Reply · See Translation · 1y


Grosse pute retourne sucer des bites Je chié dans ta gueule espèce de Salope  1

Like · Reply · See Translation · 1y

Like · Reply · See Translation · 1y

Sale racaille tu mérites de pourrir en taule ou au bout d'une corde  2

Like · Reply · See Translation · 30w

Et va te faire enculer toi et ton vaccin veran de mes couilles A flinguer toutes ces merdes  1

Like · Reply · See Translation · 23w · Edited

Like · Reply · See Translation · 1y

Examples of comments that were restored

Faut le crevé ce fils de pute



1 r.

██████████ un gnoule ? Ton cerveau il doit être bien profond dans ton fion toi .

Like · Reply · [See Translation](#) · 1y

Faut plus parler ou légiféré ,faut les buter ces chasseurs de merde !

Like · Reply · [See Translation](#) · 44w 👍 1

Toi fils de pute de LEGENDRE, quant on va te choper on va te couper les couilles et te les faire bouffer, sale mange merde.

Like · Reply · [See Translation](#) · 1y

Je prie dieu que covi infecte la France israélite et leurs alliés turcs israpoubelle et les arabes . Avec une sécheresse à tuez même Les race de vipère

Like · Reply · [See Translation](#) · 1y

Commentary

"This timely report provides important insights into the dire state of content moderation in the run up to the 2022 elections in France.

As this report finds, human oversight is necessary throughout all steps of the automated process, but we cannot rely on self-moderation by the platforms alone. We urgently need a robust framework that targets and demonetises disinformation and hate speech, coupled with enforcement for the platforms which do not do enough to take down hate speech and illegal content. Furthermore, we need targeted measures to increase the transparency around the algorithms used by platforms, which currently amplify sensationalist and false information online, and reinforce racial and gendered bias.

Much of the hate speech and disinformation we see online takes a gendered or racial dimension. **These types of disinformation not only have the power to severely harm the lives of women, people of colour and LGBTI+ people, but they also represent a clear and constant attack on the foundations of our societies.** This is part of a broader political strategy to undermine equal participation in democratic processes and to undermine our European values of gender equality, freedom and human rights.

With legislative elections approaching, not only in France, but also elsewhere in Europe, this **report highlights how much work there is still to be done.** Online content has an enormous effect on our democratic processes. When left unmoderated, it can have harmful consequences both online and offline. **We need legislation that puts democracy, human rights and privacy at its core, and for platforms to prioritise people, not profit.**

This report, compiled by HateAid and LICRA, is an important first step, but we still need more research. **Enabling civil society and academic researchers to have access to data** collected by major platforms, as has been proposed in the Digital Services Act (DSA), will allow for more research like this report. We need action now to ensure there is an intersectional approach to challenge false narratives, as well as more fact checking and financing available for such research."

Gwendoline Delbos-Corfield

Greens/EFA, Member of the European Parliament



Recommendations from HateAid and LICRA to the EU lawmakers on the Digital Services Act following the findings of the report

- I. Give all users a right to complain about wrongful content decisions made by online platforms

In the fast-paced online traffic, where 309 million people in Europe use Facebook daily⁶, it is expected that errors in the content moderation will happen. Often these errors have adverse effects on individuals and democratic events like elections and overall public discourse.

⁶ "Meta Earnings Presentation Q4, 2021", Meta 2021, https://s21.q4cdn.com/399680738/files/doc_financials/2021/q4/Q4-2021_Earnings-Presentation-Final.pdf

Users whose notices have been **rejected** by online platforms, should have **a right to a second assessment** through an internal complaint mechanism to be able to challenge wrongful platform decisions, as highlighted in the finding of this experiment.

Furthermore, wrongful content decisions are often made due to insufficient staffing of human content moderators, lack of moderator training, and/or lack of moderators who are proficient in the variety of languages used. It is important to ensure that platforms provide details of the human resources they have in place for content moderation in **a public annual report**.

II. Don't grant a free pass to online platforms to leave unlawful abuse online

In reality, what motivates the platform to delete the notified unlawful piece of content through official reporting mechanisms, be it racist hate speech or incitement to violence, is a fear of being held accountable. However, law-makers risk giving **a free pass** to online platforms to leave **unlawful content online** with no accountability. Policymakers should ensure that all notices are **thoroughly assessed** by the online platforms, without lowering the standard for assessment. Otherwise, they risk enabling a free flow of unlawful hate speech and lowering the bar for already under-resourced content moderation systems and practices, that in the case of Facebook, have already been criticised by international organisations and civil society groups for contributing to real-life violence against ethnic and religious groups in [Myanmar](#) and [India](#). The latter is the biggest market in the world where Facebook operates.

III. Provide users with an effective help-line from authorities and online platforms

Users are often left alone when dealing with online violence on social media. Victims describe a **sense of helplessness** and isolation. The current Russian invasion in Ukraine has shown the platforms' ability to react, mobilise and assign resources when under political pressure. We need a regulation that would mandate the necessary support on a day-to-day basis:

- **Enable authorities** to help users whose rights are violated by requesting platforms to remove or suspend access to the illegal content in question;
- Online platforms should **establish contact points for consumers** that should not only rely on automated means of communication, and be available in one of the official languages of each Member State.
- In order to ensure effective communication and enforcement of rules towards platforms there should be a **point of contact in every Member State**, accessible for users and authorities. This point of contact should be able to receive notifications as well as documents including those initiating proceedings against the platform in a legally binding way. This would **lower the threshold** for **victims** of online violence to **defend themselves** in front of a court.

IV. Be realistic in obligations for NGO trusted flaggers

An effective system of trusted flagging heavily relies on the civil society - often publicly or donor funded NGOs, like HateAid and LICRA, that have the best incentives to become a trusted flagger and **do not receive additional funds** for doing this job. It is important to **not overburden NGOs** with red tape, too many reporting obligations that require expensive technical equipment and human resources, as well as too strict requirements to application that may deter them from becoming a trusted flagger. Instead, we suggest **shifting the burden** of reporting requirements concerning functioning of trusted flaggers from NGOs **to online platforms**, who could easily generate this information with a help of a few clicks.

Moreover the **independence of authorities** that award trusted flagger status needs to be guaranteed and organisations that were denied the status should have access to an **appeal procedure**.

V. Establish enforceable risk assessment and mitigation

Similarly, as a car would not enter the market without certification and tests, tech companies should assess and address the systemic risks and run assessments before the products and features of their systems, including algorithms, get to users. Documents revealed by Facebook whistleblower Frances Haugen, gave an insight into the role of algorithmic amplification in spreading hate speech to drive user-engagement – with a proper risk mitigation, and **strong enforcement** in place, it should have not happened. Furthermore, the data provided by the platforms to conduct the risk assessment should be **independently verified**.

VI. Enable NGOs to do public interest research on Tech

Civil Society has been at the **forefront** of defending citizen's interests, exposing rights' breaches and demanding accountability from Tech companies for decades. We ask lawmakers to acknowledge this crucial role of the Civil Society by **widening platform data access for vetted NGOs**, associations, and not-for-profit bodies. NGOs should be given a chance to obtain the platform data of societal importance to carry out research that benefits the society.

About HateAid

HateAid gGmbH was initiated in 2018. We are the first organization in Germany to offer protection from digital violence to those affected and at the same time to support effective sanctioning of the perpetrators. Moreover, we create social awareness of the destructive effects of digital hatred on our democracy. HateAid's aim is to relieve the burden of the victims of attacks, enforce their rights, deter the perpetrators, and overall strengthen our democracy and society. As part of the Landecker Digital Justice Movement, HateAid advocates for more platform responsibility on social media.

About LICRA

The International League Against Racism and Antisemitism (LICRA) was initiated in 1927, it is an INGO that has the participatory status at the Council of Europe. LICRA is an organisation combatting racism, antisemitism, xenophobia and other forms of discrimination. LICRA is profoundly attached to the values of freedom, equality, fraternity and is promoting the ideal of universalism. Its actions are based on a network of volunteers present in Europe and especially in France. LICRA is a member of the Conference of International Non-Governmental Organisations of the Council of Europe, in which she is presiding the "Artificial Intelligence and Human Rights" committee. LICRA has been very active in the Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI) and in the Committee of Experts on Combating Hate Speech (ADI/MSI-DIS) since their creation.