Policy Brief

June 2021

Proposed EU Digital Services Act (DSA)

"notice and (NO) action": Lessons (not) learned from testing the content moderation systems of very large social media platforms

About CEP

The Counter Extremism Project (CEP) is an international, non-profit policy organization that has been engaged in efforts to effectively regulate social media and video sharing companies since 2015. Our focus lies on extremist ideologies and on illegal and terrorist content online. CEP advisors have been working with EU institutions and EU Member States for the past several years on some of the key issues the DSA aims to regulate.

Alexander Ritzmann is a Senior Advisor to CEP where he focusses on effective regulation and compliance of "social media" since 2015, working on the EU Internet Forum, the German NetzDG law, the EU Terrorist Online Directive and the DSA.

All CEP papers on the EU DSA draft can be accessed here: https://bit.ly/2RwWYrM

Key findings of independent monitoring reports

1) "notice and action" systems seem to not work properly

Based on six independent monitoring reports, the overall average takedown rate of illegal content by very large platforms (gatekeepers) based on user notices is 42%. This finding could be considered a *disprove* of concept for voluntary content moderation, because if even reported illegal content is mostly left online, the implications for legal but harmful content are obviously very negative.

2) "trusted flaggers" might be too involved with the platforms they monitor

Trusted flaggers, which are supposed to play a key role in the notice and action framework, are often underfunded and to some degree dependent on the platforms they are monitoring. There are also indications that in some cases the platforms were aware of upcoming monitoring activities which might have influenced the overall monitoring results.

Recommendations for the draft DSA

1) (Article 19) - Ensure financial independence of trusted flaggers by creating an EU wide monitoring fund which is financed by contributions of the companies falling under the DSA in proportion to their average monthly users in the EU. This EU DSA monitoring fund should be administered by the EU Commission or the European Board for Digital Services, not by the companies which would be monitored nor by EU Member States.

2) (Article 7) - Ensure the protection of EU citizens online from illegal extremist/terrorist content by mandating gatekeepers to use proactive measures with strict rules on transparency, auditability and effectiveness of the applied automated decision making systems. This approach will protect civil liberties of users much more than trusting the voluntary efforts of the companies. According to gatekeepers, the sheer amount of content forces them already to extensively apply proactive technical measures like upload- and re-upload filters to tackle illegal (or unwanted) content.

Built on sand:

The two pillars of content moderation designed by the DSA

The DSA draft aims “to create a safer digital space in which the fundamental rights of all users of digital services are protected”.[1] To achieve this objective, content moderation systems are supposed to be based on two pillars: the voluntary activities of gatekeepers and the “notice and action" mechanism for users. Unfortunately, the DSA does not address the continued failure of the existing “notice and action” moderation systems of gatekeepers that have been highlighted repeatedly in studies and tests by different organisations like CEP, jugendschutz.net, INACH and Data Intelligence Analytics. These findings could be considered a *disprove* of concept, because if not even reported illegal content is being taken down effectively, the implications for borderline content or legal but harmful content are obviously negative. Despite these alarming findings, the draft DSA (Article 14), perpetuates the “notice and action” mechanism as the main content moderation system, expecting the 400.000.000 internet users in the EU first to be exposed to illegal and possibly harmful content and then to notify the platforms about it.

The proposed content moderation approach of the DSA means the externalization of safety and security functions to users rather than a requirement for platforms to ensure the safety of their customers or prevent harmful effects on the societies in which they conduct their commercial activities.

Are trusted flaggers independent enough?

Article 19 of the DSA pushes for gatekeepers to work with “trusted flaggers”. Some of the monitoring studies indicate indeed that illegal content reported by trusted flaggers sometimes has a higher chance of being taken down or blocked than if the exact same illegal content is reported by those flaggers pretending to be simple users. Unfortunately, the DSA misses out on solving significant problems that come with the concept of trusted flaggers, for example:

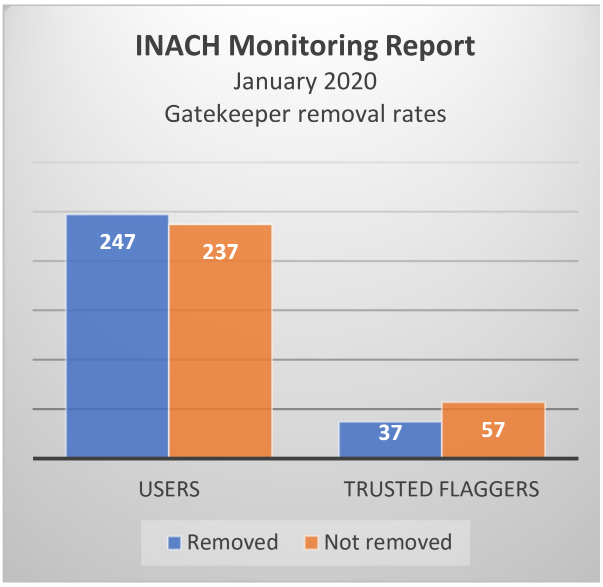
1) At the moment, many trusted flagger organizations are small civil society organizations that provide their expertise in the framework of small and short-term projects. Those trusted flaggers mostly manually search for illegal content and are by no means capable of monitoring the millions of pieces of content that are uploaded each day on the very large platforms.

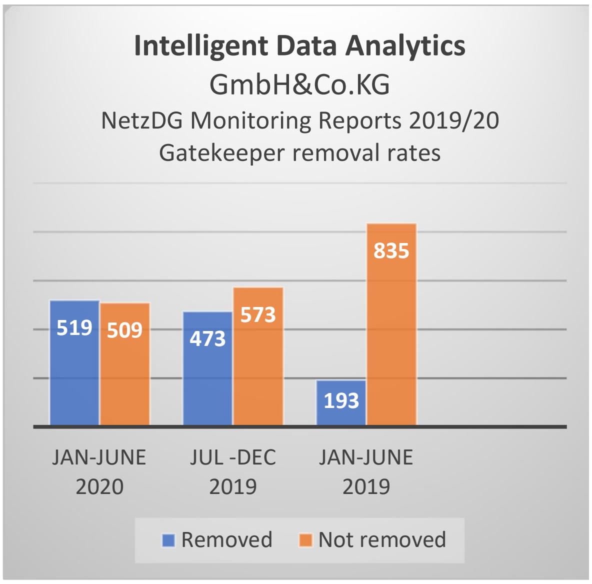
2) As highlighted by a recent report published by the Council of Europe, the gatekeepers are expected to simultaneously support the trusted flagger organisations they work with and to respect their independence. The report states: “In practice it is typically Internet platforms that approach organisations about becoming trusted flaggers, and it is in the gift of Internet platforms to offer or withhold that status. In one sense this is a relationship in which nearly all the power lies with the Internet platform“.[2]

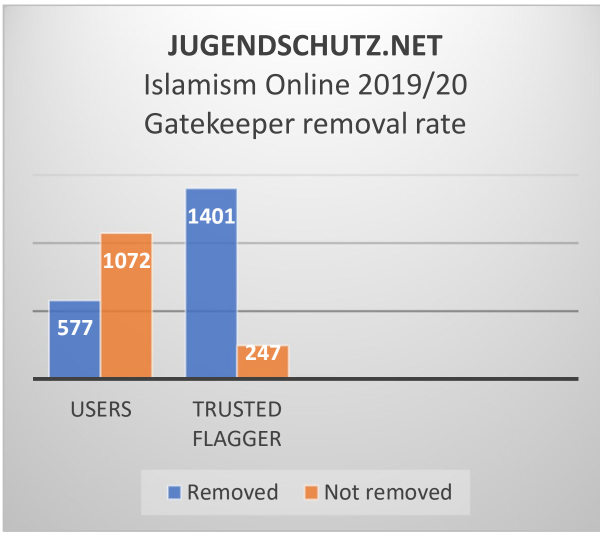
3) The same report points out this important point: “Internet platforms already have a very close working relationship with trusted flaggers, this could make it easier for the platforms to gain information, directly or indirectly, about when the monitoring period is underway.“ This concern is being shared by some trusted flagger organisations themselves.[3]

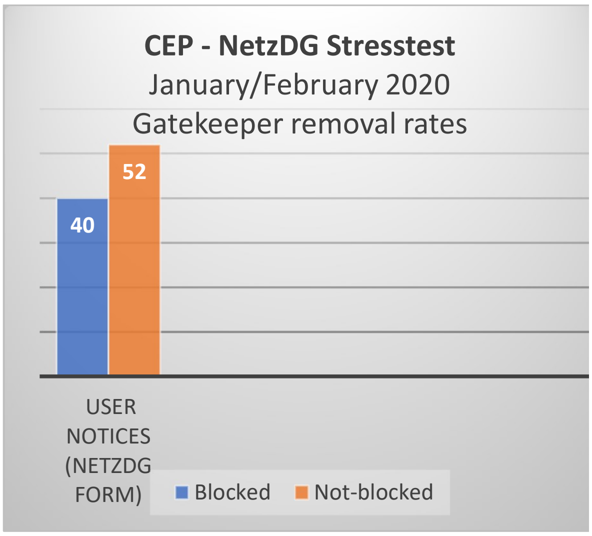
Overview of the monitoring reports

In 2019 and 2020, several monitoring activities were conducted by Jugendschutz[4], INACH[5], Intelligent Data Analytics[6] and CEP[7]. Based on the analysis of six monitoring reports, the **overall average takedown rate of illegal content** by gatekeepers on their platforms based on **user notices is 42 %,** for **trusted flagger notices it is 62 %.** Naturally, the findings fluctuate depending on the platforms, the country the monitoring took place in and time of the monitoring. However, the overall implications are clear: Content moderation systems that rely on notice and action, as proposed in the draft DSA, cannot provide the promised and necessary safety for EU internet users.









[1] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

[2] https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d

[3] Interviews with staffers of two trusted flagger organisations in March and May 2021.

[4] https://www.jugendschutz.net/fileadmin/download/pdf/Bericht_2019_2020_Islamismus_im_Netz.pdf

[5] https://www.inach.net/wp-content/uploads/INACH_Monitoring-Report.pdf

[6] https://www.carlgrossmann.com/liesching-das-netzdg-in-der-praktischen-anwendung/ (Page 40)

[7] https://bit.ly/2SmCRwO