How to protect fundamental rights when appointing Trusted Flaggers

A guide for regulators

Recommendations

DSC must ensure a reasonable balance of Trusted Flagger (TF) entities. Considering the restriction of the numbers of TF under recital 62 DSA, it should be ensured: First, that the lack of resources does not result in excluding smaller NGOs from becoming TF, particularly those working with marginalised groups. The broad reporting obligations are a particular burden to smaller NGOs or civil society organisations. Second, that TF are not designated on a first come first serve basis but based on merit and expertise.

DSC should agree to not award TF status to right holders. The industry has long history of false copyright claims that result in the systematic over removal of entirely legal user content. The industry's inherent commercial incentives for the quick, automated removal of user content in combination with low to no costs for wrongful decisions, makes right holders and their representatives unfit to act as Trusted Flaggers.

DSC should agree to not award TF status to law enforcement agencies as it effectively turns police authorities into judges and blurs the democratic firewall between the executive and the judicial branches. Giving law enforcement agencies TF powers risks diverting policing resources away from actual penal action and undermines several important fundamental rights, in particular the right to due process and a fair trial as well as the presumption of innocence until a suspect is found guilty by a court.

1. Purpose of this guide

The Digital Services Act (DSA) aims to regulate online platforms and safeguard fundamental rights in digital contexts. One crucial aspect of the DSA is the incorporation of Trusted Flaggers in its enforcement ecosystem. These are entities that have a particular expertise in tackling illegal content online.

Trusted Flaggers have great potential, but they can also be used as a vehicle for specific interest groups or governments to obtain an outsized or even illegitimate influence by means of over- or under-blocking.[1] Consequently, the trusted flagging arrangements can reflect and reinforce pre-existing power structures, including state coercion and private power.[2] It is therefore crucial to establish best practices for designating Trusted Flaggers in line with human rights.

This guide develops best practices for designating Trusted Flaggers by considering (1) the stakeholders involved in their designation, (2) users' rights, and (3) the potential impact on freedom of expression and online content moderation. Importantly, the guide acknowledges the disproportionate impact of online violence in marginalised communities such as women, girls, and Black Indigenous People of Colour (BIPOC).

These are the key questions that will be addressed:

1. How can Digital Services Coordinators (DSCs) ensure the expertise, independence, and objectivity of Trusted Flaggers, and what standards should be considered to fulfil these criteria effectively?
2. How can DSCs ensure Trusted Flaggers and online platform providers work together in compliance with the EU Charter of Fundamental Rights?
3. How can DSCs ensure transparency and prevent conflicts of interest of Trusted Flaggers?

2. What are Trusted Flaggers?

Trusted flaggers are entities with particular expertise in tackling or flagging illegal or harmful content online, and that have been awarded an official status to do so under the DSA. Flagging' is the process by which third parties can report content to platforms for content moderation review. Since Trusted Flaggers acquire certain privileges in flagging, their role is debated: By some, Trusted Flaggers are seen as a solution to the problems which the platforms themselves lack the incentives, expertise, responsibility or legitimacy to solve.[3] As such, they are perceived as a source of trustworthy expertise and representation.

1 Dvoskin, Brenda, "Representation without Elections: Civil Society. Participation as a Remedy for the Democratic Deficits of Online Speech Governance", Villanova Law Review (15 December 2021). Available at https://ssrn.com/abstract=3986181.

2 Schwemer, Sebastian Felix, "Trusted Notifiers and the Privatization of Online Enforcement", 20 November 2018, Computer Law & Security Review, Volume 35, Issue 6. Available at https://ssrn.com/abstract=3287754.

3 Appelman, Naomi & Leerssen Paddy, "On Trusted Flaggers", 2022. Available at https://law.yale.edu/sites/default/files/area/ center/isp/documents/trustedflaggers_ispessayseries_2022.pdf.

By others, Trusted Flaggers are critically seen as an unaccountable co-optation by public and private power,[4] that can be used as a vehicle for specific interest groups or governments to obtain an outsized or even illegitimate influence over the free expression of billions of people online. Most significantly, concerns are raised about the lack of transparency, accountability, and contestability of content moderation practices.[5] Consequently, the trusted flagging arrangements can reflect and reinforce pre-existing power structures, including State coercion and private power.

In summary, Trusted Flaggers can be seen as tools to make content moderation more effective and legitimate, but also as self-interested co-opters spurring its worst excesses. In practice – regardless of the perspective – the Trusted Flagger mechanism means outsourcing knowledge and decentralising some of the control over content moderation on Big Tech platforms.

3. Trusted Flaggers in the Digital Services Act

While the Digital Services Act does not specifically define Trusted Flaggers, the European Commission provides an insight. According to the EU Recommendation on measures to effectively tackle illegal content online (2018), a Trusted Flagger is "an individual or entity which is considered by a hosting service provider to have particular expertise and responsibilities for the purposes of tackling illegal content online." Similarly, recital 61 of the DSA clarifies that Trusted Flaggers are entities or organisations. This can be either public entities or non-governmental organisations, but explicitly not individuals. Therefore, the nature and mission of Trusted Flaggers can vary significantly.

Under Art. 22(2) DSA, the status of a Trusted Flagger can be awarded upon application. To qualify, applicants must verifiably meet several conditions. The designation will be done by the Digital Services Coordinator (DSC) of the Member State in which the applicant is established.

Firstly, Trusted Flaggers must have expertise and competence for the purposes of detecting, identifying, and notifying illegal content. Secondly, they must represent collective interests and be independent from any online platform. Thirdly, they must perform their activity for the purpose of submitting notices in a timely, diligent, and objective manner. Finally, they must have transparent funding structures and need to publish at least once a year a report on actions and notices made in the previous year, which must meet certain minimum requirements.

In terms of maintenance and length of the awarded position, authorised Trusted Flaggers are listed and regularly updated by the Commission in a publicly accessible database (cf. Article 19(4) DSA). If the entity no longer meets the necessary legal requirements (statutory

4 Schwemer, Sebastian Felix, "Trusted Notifiers and the Privatization of Online Enforcement", 20 November 2018, Computer Law & Security Review, Volume 35, Issue 6. Available at https://ssrn.com/abstract=3287754.

5 Bloch-Wehba, Hannah, "Global Platform Governance: Private Power in the Shadow of the State", 10 May 2019, SMU Law Review, Vol. 72, 2019, Texas A&M University School of Law Legal Studies Research Paper No. 20-18. Available at SSRN: https://ssrn.com/abstract=3247372.

conditions), the DSC may revoke the status pursuant to Art. 22(7) DSA following an appropriate investigation. This investigation is either initiated independently or based on information received from third parties or an online platform.

4. How Trusted Flaggers can be useful

Legitimacy and better understanding of the different contexts: Trusted Flaggers can represent marginalised or under-represented communities that are disproportionately impacted by online violence and hate speech. One example is the case of marginalised groups as they face difficulties in obtaining adequate protection and redress from platforms like “race-blind” content moderation policies.[6] In this context, the Trusted Flagger model would allow representative interest groups to acquire a stake in platform moderation and give a voice to voiceless parties.

The process of transferring notifications about illegal content to humans for moderation is often slow, leaving victims of online violence unprotected. The Trusted Flagger model should allow a fast track of flagged content by these entities to ensure more expedited reaction and moderation. This is particularly important when harmful trends online are identified targeting marginalised groups as the trend emerges and before it becomes viral, thus reducing the dissemination of illegal content or content that violates a platforms terms of service.[7] The process has the potential to provide better local context for escalation processes of violence, especially during important events such as elections.[8]

Users not often have the means or knowledge on how to proceed in cases of content moderation or are left with uncertainties in front of the slow reaction from online platforms. While the DSA provides new mechanisms that make user rights easier to be exercised, Trusted Flaggers can provide an extra layer of support and protection. This is particularly important for having the perspective of victims of online violence better represented. Victims that have not had a timely reaction from online platforms can reach out to Trusted Flaggers for quicker remedy. For them, the removal of content that contains death threats, doxxing or defamation needs to be effective urgently to minimise the potential damage to their reputation, harming their families and negatively affecting their mental health.[9]

6 Ángel Díaz & Laura Hecht-Felella, "Double Standards in Social Media Content Moderation", Brennan Ctr. for Just. (4 August 2021). Available at https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation. Caitlin Ring Carlson & Hayley Rousselle, “Report and Repeat: Investigating Facebook’s Hate Speech Removal Process”, 2020, Available at https://journals.uic.edu/ojs/ index.php/fm/article/view/5<>66. Julia Angwin & Hannes Grassegger, “Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children”, ProPublica (28 June 2017). Available at https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

7 National Democratic Institute, “Interventions for ending online violence against women in politics”, 2022. Available at https://www.ndi.org/sites/default/files/NDI%20Interventions%20to%20End%20OVAW-P.pdf.

8 Idem.

9 Sincek, Daniela; Duvnjak, Ivana; Milić, Marija, “Psychological Outcomes of Cyber-Violence on Victims, Perpetrators and Perpetrators/Victims”, 2017. Available at https://repozitorij.ffos.hr/islandora/object/ffos%3A3939/datastream/FILE0/view. Sameer Hinduja, Justin W. Patchin, “Offline Consequences of Online Victimization School Violence and Delinquency”. Available at https://www.tandfonline.com/doi/epdf/10.1300/ J202v06n03_06?needAccess=true.

Even though Article 22(1) DSA specifies that the role of Trusted Flaggers is restricted to illegal content, the application of the terms of service is more likely to be used for removing content.[10] This is, based on the German experience with NetzDG, the most common explanation in content moderation for platforms and more favourable for the platforms. From all NetzDG reports, over 90% are taken down as a violation of terms of service, a legal analysis of the NetzDG is only the second step of analysis for content moderation.[11]

5. How to protect the Trusted Flagger mechanism against abuse

The DSA creates a new legal obligation for providers of online platforms to "ensure that notices submitted by Trusted Flaggers […] are given priority and are processed and decided upon without undue delay" (Article 22 DSA). The additional processing speed required for notices submitted by Trusted Flaggers is meant to ensure that victims of illegal online content such as threats of violence are protected as quickly as possible from harm.

Any legal requirement to be quick, however, also creates the risk that the platforms concerned are less thorough in their assessment as to whether any user-generated content that a Trusted Flagger alleges to be illegal is, in fact, illegal. There is ample[12] (and sometimes sadly ironic[13]) evidence[14] from jurisdictions in which Trusted Flaggers and large online platforms have maintained close relationships that legal pressure to act quickly often leads to a dramatic decrease in the accuracy of illegality claims, as well as an increase in wrongful content take-downs.

5.1. Right holders as Trusted Flaggers

In the U.S. many online platforms automatically accept automated take-down requests without further verification if they come from Trusted Flaggers. Some provided no human review at all for most automated notices they received, and instead relied on direct back-end take-down privileges for trusted copyright holders.[15] As a result, researchers documented

10 Friehe, Matthias, "Löschen und Sperren in sozialen Netzwerken", Neue Juristische Wochenschrift 2020, 1697-1760.

11 Facebook, NetzDG Transparency Report (January 2022). Available at https://about.fb.com/de/wp-content/uploads/sites/10/2022/01/NetzDG-EN.pdf.

12 Daphne Keller, Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List, Stanford Center for Internet and Society, February 8, 2021. Available at https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

13 Neely, Adam: Warner music claimed my video for defending their copyright in a lawsuit they lost the copyright for, 6 February 2020. Available at https://www.youtube.com/watch?v=KM6X2MEl7R8.

14 Urban, Jennifer M. and Karaganis, Joe and Schofield, Brianna and Schofield, Brianna, Notice and Takedown in Everyday Practice (March 22, 2017). UC Berkeley Public Law Research Paper No. 2755628, Available at https://ssrn.com/abstract=2755628.

15 Idem.

considerable error in U.S. DMCA[16] take-downs: Among notices submitted to Google Search, for instance, they found questionable legal claims in 28 percent of the cases.

While the DSA's recitals provide examples for what kind of organisations can be awarded the Trusted Flagger status, based on the existing empirical evidence it is essential that this privilege is not given to entities with a vested interest in the removal of user-generated online content to prevent the system's abuse. Platforms should also be strongly discouraged from blindly trusting content notices without human review simply because they come from a Trusted Flagger.

In the DSA's recitals, legislators have included some examples as to who they expect to be eligible to apply for Trusted Flagger status: Trusted flagger candidates "can be public in nature, such as, [...] internet referral units of national law enforcement authorities or of the European Union Agency for Law Enforcement Cooperation ('Europol') or they can be non-governmental organisations and private or semi-public bodies such as the organisations part of the INHOPE network of hotlines. […] In particular, industry associations representing their members' interests are encouraged to apply for the status of Trusted Flaggers" (Recital 61, DSA).

However, it is important to note that this recital does not create an obligation for Digital Services Coordinators to award Trusted Flagger status to any such group or organisation, should it consider that the requirements set forth by the DSA are not fulfilled. As discussed above, these requirements include, for instance, the expertise for detecting, identifying and notifying illegal content, the independence from any online platform, as well as the ability to submit notices in a timely, diligent and objective manner.

While industry associations that represent rights holders might likely be able to demonstrate the required expertise in detecting, identifying and notifying user content that violates their copyright, it appears questionable whether those associations are also able to perform those tasks with the necessary diligence and objectivity. As the examples above show, rights holders have a long history of false copyright claims that result in the systematic over removal of entirely legal user content; a conduct that makes sense for rights holders from a commercial perspective. To increase the flagging system's efficiency, industry associations use automated detection to identify infringing content. However, those automated systems are often unable by design to distinguish between a permitted use of, say, a specific melody or piece of music, and an illicit one. What is more, automated detection systems are inadequate in telling the difference between Bach's "Goldberg Variations" recorded by star pianist Lang Lang and the same piece played by a random user in their living room. The copyright of the first is owned by Deutsche Grammophon GmbH, the other might be published under a Creative Commons license.

According to an empirical study of content take-downs, "what is alarming is the magnitude, frequency and systematic nature of these errors, which remained undetected and uncorrected for months on end. While we may excuse these errors on the basis that they arose from programs that are misconfigured with wrong information, automated systems propagated

16 The Digital Millennium Copyright Act (DMCA) is a U.S. copyright law from 1998 that obliges online platforms to promptly block access to or remove alleged infringing material when they receive notification of an infringement claim from a copyright holder or the copyright holder's agent.

these errors across hundreds and thousands of takedown requests".[17] The researchers not only attributed this to machine errors alone, but linked the wrongful removal of legal user content directly to the lack of diligence and objectivity of the flagger: "If the price of each arrow is low or minimal, to improve his chances, the reporter will fire off as many arrows as he could to hit a target, regardless of the accuracy or precision".[18]

5.2. Law enforcement as Trusted Flaggers

The DSA recitals also mention national law enforcement and Europol as candidates to become Trusted Flaggers. While it can be expected that law enforcement authorities should have expertise in detecting, identifying and notifying illegal content, as well as sufficient independence from the online platform hosting said content, giving the police direct and privileged removal powers of people's online speech would constitute a shift of focus: away from prosecuting persons suspected of an illegal activity and towards simply making that activity invisible on social media platforms.

While police forces are required "to support the judiciary in bringing offenders to justice"[19] in cases of a breach of the law, the overemphasis of and privileged access to online content removal tools is likely going to divert resources away from actual prosecution. For example, in cases of minor alleged offences, or where the identification of the offender would require additional resources from police often already operating under severe resource constraints.

As a result, instead of following due process when encountering potentially illegal or even criminal behaviour online, and guaranteeing a fair trial in view of an independent judgement in court, police Trusted Flaggers are incentivised—not least by strained resources—to decide the legality of the online content on their own, thus moving the power to assess the legality from the judiciary to the executive branch of government.

The Santa Clara Principles on Transparency and Accountability in Content Moderation, one of the most complete sets of principles in this space, specifically scrutinise the involvement of government in content moderation: "Companies should recognise the particular risks to users' rights that result from state involvement in content moderation processes. This includes a state's involvement in the development and enforcement of the company's rules and policies, either to comply with local law or serve other state interests. Special concerns are raised by demands and requests from state actors (including government bodies, regulatory authorities, law enforcement agencies and courts) for the removal of content or the suspension of accounts".[20]

In theory, online platforms are of course not required to remove user content notified by law enforcement authorities or other state institutions that are awarded the Trusted Flaggers status. In practice, however, online platforms are "inherently biased in favor of the govern-

17 Daniel Seng, "Who Watches the Watchmen? An Empirical Analysis of Errors in DMCA Takedown Notices", SSRN Research Paper no 25632023, 3. Available at https://ssrn.com/abstract=25632023.

18 Idem.

19 OSCE, "Guidebook on Democratic Policing by the Senior Police Adviser to the OSCE Secretary General", 2nd Edition (May 2008). Available at https://www.osce.org/files/f/documents/5/3/23804.pdf.

20 The Santa Clara Principles On Transparency and Accountability in Content Moderation (2020). Available at https://santaclaraprinciples.com.

ment's favored positions".[21] Except for very mediatised cases like the San Bernardino shooting, where Apple refused to assist the FBI in unlocking the dead shooter's iPhone, online platforms have little incentive to object to law enforcement requests for user content removal requests, and even less so once a police force's Trusted Flagger status compels platforms to treat such notices with particular speed. Unlike in the San Bernardino case, even wrongful content removals entail mostly low or no costs for online platforms, whereas objecting to a flagging by law enforcement may expose providers to serious public and political relations risks.

The resulting incentive structure makes it "easy for government—and particularly law enforcement—to use the [Trusted Flagger] systems to coerce and pressure platforms to moderate speech they may not otherwise have chosen to moderate,"[22] despite the absence of any legal requirement to comply with flaggings.

Giving law enforcement the status of Trusted Flaggers under the DSA therefore risks effectively turning police authorities into judges and blurring the democratic firewall between the executive and the judicial branches. Acting in this way undermines several important fundamental rights, in particular the right to due process and a fair trial as well as the presumption of innocence until a suspect is found guilty by a court.

6. The role of online platforms

The role and responsible behaviour of online platform providers will be a key element for the functioning of the Trusted Flagger model. For this aim, online platforms should follow best practices that ensure a fair, transparent, and consistent application of their moderation policies.

In the past, VLOPs have built relationships with civil society, however, this has been mainly ad-hoc and left stakeholders often without follow-up or acknowledgement about whether their recommendations or flagging have been considered or implemented.[23] While legislation will make sure that VLOPs are held accountable, fostering meaningful relationships with their teams can make the processes less laborious and reduce the burden on Trusted Flaggers, especially the ones that have less resources. Still, these relationships are not reliable for civil society organisations, as they depend completely on the platform's goodwill to cooperate. For example when contact points are no longer available after civil society has published critical reports or filed strategic lawsuits with publicity.

However, as important as this process is, it is essential that relationships do not interfere with the independence of either party. Instead, they should improve the predictability of content moderation decisions for Trusted Flaggers and concerned users.

21 David Greene, Paige Collings, and Christoph Schmon, "Online Platforms Should Stop Partnering with Government Agencies to Remove Content", EFF (12 August 2022). Available at https://www.eff.org/deeplinks/ 2022/08/online-platforms-should-stop-partnering-government-agencies-remove-content.

22 Idem.

23 Dvoskin, Brenda, "Representation without Elections: Civil Society. Participation as a Remedy for the Democratic Deficits of Online Speech Governance", Villanova Law Review (15 December 2021). Available at https://ssrn.com/abstract=3986181.

Algorithmic content moderation often fails to respect human rights standards in its decision-making. Platforms should ensure to invest sufficient resources in human content moderation, so that the option to transfer potentially illegal content flagged by Trusted Flaggers to human staff is always available.

Online platforms should also encourage the creation, contribution to, and use of a shared repository of localised lexicons with examples of hate speech and other infringing online content. While Trusted Flaggers can make the reporting and flagging more accurate based on their expertise, VLOPs should use that knowledge to continuously grow their understanding of hate speech and online violence in the diverse cultural and linguistic contexts.[24]

While the DSA provides new redress and complaint mechanisms, online platforms should ensure that user content that is being flagged by third parties, including Trusted Flaggers, is treated according to strictly and publicly documented moderation policies and processes. VLOPs should also increase transparency and accountability by making their moderation policies and processes available to the public, including all rules and policies given to platforms' moderation teams. If processes are not well documented and decisions made transparent, the Trusted Flagger mechanism could even reduce the accountability of the organisations involved.[25]

****

24 National Democratic Institute, "Interventions for ending online violence against women in politics", 2022. Available at https://www.ndi.org/sites/default/files/NDI%20Interventions%20to%20End%20OVAW-P.pdf.

25 Seng, Daniel Kiat Boon, "The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices" (31 January 2014). 18 Va. J. L. & Tech. 369. Available at https://ssrn.com/abstract=2411915.