Trusted Flaggers

Appelman, N.M.I.D.; Leerssen, P.J.

Publication date
2022
Document Version
Final published version

On "Trusted" Flaggers

Naomi Appelman & Paddy Leerssen

introduction

Trusted flaggers are on the rise in platform governance. Platforms are entering into a growing array of trusted flagging arrangements – also referred to as trusted 'notifiers', 'reporters, 'partners', and so forth. The concept has also recently started appearing in legislation. And yet, the meaning of this concept remains vague and contested. Flagging' is the process by which third parties can report content to platforms for content moderation review. By now a "ubiquitous mechanism of governance", flagging is in principle open for all to use.[1] But some flaggers are more equal than others. We introduce a concept of "trusted flaggers" that describes, broadly speaking, how third parties have acquired certain privileges in flagging. The privileges to the trusted third party typically include some degree of priority in the processing of notices, as well as access to special interfaces or points of contact to submit their flags.

Trusted flagging complicates an already-controversial process. Even more so than conventional flagging,[2] trusted flagging outsources part of the responsibility for content moderation from platforms to third parties.[3] This diffusion of responsibility is precisely what makes trusted flagging both attractive and controversial. It is often cited as a solution for which platforms themselves lack the incentives, expertise, or legitimacy.[4] But

[1] Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Soc'y 410 (2016).

[2] *Id*.

[3] Sebastian Schwemer, *Trusted Notifiers, and the Privatization of Online Enforcement*, 35 Comput. L. & Sec. Rev. 105339 (2019).

[4] *E.g.*, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Illegal

trusted flagging is also seen as a vehicle for specific interest groups or governments to obtain an outsized or even illegitimate influence by mean of over-blocking.[5] In short, not everyone trusts the same flaggers.

This essay unpacks the practices of trusted flagging. We first discuss self-regulatory flagging partnerships on several major platforms. We then review various forms of government involvement and regulation, focusing especially on the EU context, where law-making on this issue is especially prevalent. On this basis, we conceptualize different variants of trusted flagging, in terms of their legal construction, position in the content moderation process and the nature of their flagging privileges. We then discuss competing narratives about the role of trusted flaggers; as a source of expertise and representation; as an unaccountable co-optation by public and private power; and as a performance of inclusion. In this way, we illustrate how "trusted flagging," in its everyday operationalization and critique, serves as a site of contestation between competing interests and legitimacy claims in platform governance.

1. Varieties of trusted flagging: an overview of practices

1.1 Platform policies

Platform policies on trusted flagging are a useful starting point for our discussion. Some platforms explicitly mention "trusted flaggers," or its variants, but there is also a broader ecosystem of institutions and practices with similar functions referred to by other names.[6] This essay will try to account for both. Platforms often grant flagging privileges voluntarily as part of their own content moderation policies. Importantly, these private ordering constructs follow quite naturally from conditional liability regimes such as in the EU's eCommerce Directive or the US' Digital Millennium Copyright Act ("DMCA"), which attach binding liabilities to

Content Online: Towards an enhanced responsibility of online platforms, COM (2017) 555 final (Sept. 28, 2017) [hereinafter Communication from the Commission].

[5] *E.g.*, *EU Fails to Protect Free Speech Online, Again*, Article 19 (October 5, 2017), https://www.article19.org/resources/eu-fails-to-protect-free-speech-online-again.

[6] For instance, YouTube only refers to "Trusted Flaggers" program, whereas Facebook refers to both trusted flaggers and "Trusted Partners," which perform a comparable function. TikTok calls them "Safety Partners," and Twitter refers to "Trusted Partners." Of course, the practices described here also differ somewhat in scope. For example, in TikTok's case, "Safety Partners" at times also refers to NGOs and other entities that advise on the drafting of community guidelines.

certain types of flagging.[7] However, forms of (regulated) self-regulation as well as co-regulation are on the rise, as will be discussed in section 1.2.

Diving into platform policy, YouTube is a good place to start since it has a dedicated and easily accessible policy on trusted flaggers.[8] For YouTube, trusted flagger status is open to individuals as well as government agencies and NGOs that have proven expertise in one of the "policy verticals" in their community guidelines.[9] These flaggers are granted "prioritized flag reviews for increased actionability" as well as access to "[a] bulk-flagging tool that allows for reporting multiple videos at one time; visibility into decisions on flagged content; [and] ongoing discussion and feedback on various YouTube content areas."[10] Importantly, the program is only directed at content flagged for community guidelines violations, not violations of national legal norms. Reports by a trusted flagger are reviewed with priority but otherwise still subject to the normal review process. Notably, all trusted flaggers are subject to a non-disclosure agreement, and there is no information available on the number of trusted flaggers participating in the program; the participating organizations; the amount of content they flag; and, finally, the percentage of content flagged by them that is actually removed. This lack of transparency will be a recurring theme across other services, too.

Twitter and TikTok do mention policies on, respectively, "Trusted Partners" and "Safety Partners," but they are not documented in the same detail as YouTube's policy. TikTok has several specific programs where they work with safety partners, such as fact-checking and media literacy.[11] Even though TikTok regularly refers to "partnerships" with many NGOs,

---

[7] *See* Directive 2000/31/EC, 2000 O.J. (L 178) 1; Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (1998). In short, both the e-Commerce Directive and the U.S. DMCA create a conditional exemption of liability where hosting providers are not liable for platforming third party speech as long as they do not have knowledge of its unlawful nature. Crucially, a notification can lead to knowledge and as such trigger liability. *See also* Aleksandra Kuczerawy, Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards (2018). On the U.S. framework, *see* James Grimmelman, Patterns of Information Law: Intellectual Property Done Right (2017).

[8] *See* About the YouTube Trusted Flagger Program, YouTube, https://support.google.com/youtube/answer/7554338?hl=en#zippy= (last visited Mar. 13, 2022).

[9] Examples of such "policy verticals" include child safety or glorifying violence. Community Guidelines, YouTube, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#community-guidelines (last visited Mar. 13, 2022).

[10] *See* About the YouTube Trusted Flagger Program, *supra* note 7.

[11] Safety Partners, TikTok, https://www.tiktok.com/safety/en-us/safety-partners (last visited Mar. 13, 2022).

it remains unclear what exactly these partnerships entail.[12] As such, it is unclear whether and which organizations have trusted flagger status. However, in September 2020, TikTok joined the EU code of conduct on countering illegal hate speech online (Code of Conduct).[13] Signatories explicitly commit to enabling civil society organizations to perform the role of "trusted reporters," and from the most recent monitoring report, it seems that TikTok indeed did so.[14]

Similarly, Twitter refers to "trusted partners" in the context of its Trust and Safety Council, which advises on content moderation issues.[15] It mentions "trusted reporters" only in its transparency reports to refer specifically to hate speech reporters in the context of the Code of Conduct.[16] As such, it is unclear to what extent Twitter and TikTok use trusted flaggers outside the context of the Code of Conduct. Finally, Facebook, similar to TikTok and Twitter, does not have a clearly outlined policy specifically for trusted flaggers, but it is a signatory to the Code of Conduct and has, therefore, a trusted reporters program dedicated to hate speech. Furthermore, Facebook mentions its trusted flaggers program at various places such as in responses to government consultations.[17]

Overall, these policies provide only a surface-level view of flagging arrangements. Most platforms do not have dedicated policies, and even if they do these are scant on details and omit many similar arrangements that go by other names such as will be discussed in the following section.

---

[12] Community Guidelines Enforcement Report, TikTok (Sept. 22, 2020), https://www.tiktok.com/safety/resources/transparency-report-2020-1?lang=en.

[13] Cormac Keenan, *TikTok Joins the Code of Conduct on Countering Illegal Hate Speech Online*, TikTok (Sept. 8, 2020), https://newsroom.tiktok.com/en-gb/tiktok-joins-the-code-of-conduct-on-countering-illegal-hate-speech-online.

[14] EU Code of Conduct Against Illegal Hate Speech Online: Results Remain Positive But Progress Slows Down, Eur. Comm'n (Oct. 7, 2021), https://ec.europa.eu/commission/presscorner/detail/nl/ip_21_5082.

[15] Our Continued Collaboration with Trusted Partners, Twitter (Dec. 17, 2021), https://blog.twitter.com/en_us/topics/company/2021/our-continued-collaboration-with-trusted-partners.

[16] Removal Requests, Twitter, https://transparency.twitter.com/en/reports/removal-requests.html#2021-jan-jun (last visited Mar. 13, 2022).

[17] Facebook Response to EC Public Consultation on the Digital Services Act (DSA), Facebook (Sept. 8, 2020), https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwi39ob8-6T2AhUgQ_EDHRoWBCAQFnoECAIQAQ&url=https%3A%2F%2Fabout.fb.com%2Fde%2Fwp-content%2Fuploads%2Fsites%2F10%2F2020%2F09%2FFINAL-FB-Response-to-DSA-Consultations.pdf&usg=AOvVaw0nfA2tiyb5XyqwGXt8UAQo; *see also* Keenan, *supra* note 13.

1.2 (Self-)Regulated Flagging

Looking beyond the labels placed by platforms, there are many arrangements that fulfil a very similar role to trusted flaggers. Even though trusted flaggers are technically in the domain of private platform policy, or at most self-regulation, the law and government agencies are involved in some capacity in many cases—either directly by government agencies submitting their own flags, or indirectly by facilitating special treatment for private flaggers. We focus on examples from the European Union, where such policies are especially common. These examples are indicative of the variety of possible legal constructions for trusted flagging.

The EU Code of Conduct on Hate Speech

The EU's Code of Conduct was a relatively early attempt to formalize flagging relationships. Organized under the auspices of the European Commission in 2016, this Code was originally signed by Facebook, Microsoft, Twitter, and YouTube, with subsequent sign-ons by Instagram, Snapchat, Dailymotion, Jeuxvideo.com, TikTok, and LinkedIn.[18] In the Code of Conduct, signatories commit to, amongst other actions, setting up trusted flagger programs and providing adequate training for civil society organizations to fulfil this role.

As discussed, the platforms' own policy documents do not clarify what steps have been taken as part of the Code, but periodical reporting for the Code from the European Commission does offer some insights. The latest report, covering 2021, shows that "trusted reporters" are relatively influential in the overall program: out of a total of 4,353 notices submitted to signatory services in that year, 1,306 were submitted through specific channels available only to trusted flaggers and reporters[19] (In absolute terms, of course, these figures pale in comparison to the billions of items moderated through the platforms' own processes). Notices submitted by trusted reports were only slightly more likely to trigger removal than those from ordinary users. Quite frequently, the trusted reporters reported

[18] The EU Code of Conduct on Countering Illegal Hate Speech Online, Eur. Comm'n (June 30, 2016), https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

[19] Didier Reynders, *Countering Illegal Hate Speech Online: 6th Evaluation of the Code of Conduct*, Eur. Comm'n (Oct. 7, 2021), https://ec.europa.eu/info/sites/default/files/factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en_1.pdf.

these same cases not only to the platform but to the police or other national authorities (a total of 315 notices).

Regulated flagging in the NetzDG and Digital Services Act

On a national level, an early mover in attempting regulation for trusted flaggers was the German NetzDG, or Netzwerkdurchsetzungsgesetz in full, which institutes a legally binding takedown procedure for content violating German law.[20] This framework allows NGOs and other third parties to submit notices on behalf of public interests, such as removing hate speech. Platforms regulated under this framework must state in their public reports how many of the complaints they received originated from these third parties (Beschwerdenstellen). For YouTube, these complaints account for more than a third of all notices received (92.424 out of 263.653).[21] Yet, on other platforms third parties are far less active: their complaints accounted for only 11% of Twitter's NetzDG flags (7,872 out of 67,950) and as little as 3% on Facebook (3666 / 111419).[22] Notably, beyond transparency, the NetzDG does little else to privilege reporting agencies or otherwise formalize their role, such as by bestowing them with priority treatment rights or special data access Interestingly, each of these platforms' reports emphasizes that they rely on self-identification by the submitting agency. For example, Google states, "We cannot verify whether a user who selects 'reporting agency' is indeed affiliated with a reporting agency."[23] Overall, the reporting agencies appear to act at an arm's length from platforms, compared to most trusted flaggers: their claims do not receive priority treatment; they do not have access to special flagging interfaces; and they are not formally recognized or accredited by the platforms.

---

[20] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken In the version of 1 September 2017 (Federal Law Gazette I, p. 3352 ff. Valid as from 1 October 2017).

[21] Removals Under the Network Enforcement Law, YouTube, https://transparencyreport.google.com/netzdg/youtube?hl=en (last visited Mar. 20, 2022).

[22] Germany, Twitter, https://transparency.twitter.com/en/reports/countries/de.html (last visited Mar. 20, 2022); Network Enforcement Act ("NetzDG"), Facebook, https://www.facebook.com/help/285230728652028

[23] Removals under the Network Enforcement Law, Google, https://transparencyreport.google.com/netzdg/youtube?hl=en (last visited Mar. 20, 2022).

The DSA represents a more robust version of regulated flagging.[24] Article 19 on "Trusted Flaggers" allows national authorities to publicly appoint organizations to submit notices that must be processed with priority.[25] Its scope is broader than the NetzDG, since these notices can also pertain to violations of platforms' own content rules. Notably, independence from other interests is not required, opening the door to trusted flagging for commercial purposes, such as IP enforcement.[26] The DSA also codifies already existing practices of publicly-appointed trusted flaggers, with specific reference to the so-called Internet Referral Units (IRUs) of national police forces which will be discussed in depth below.[27] For now, it suffices to observe that the DSA will likely formalizes, accelerate, and expand these practices across the European Union. Notably, neither the NetzDG and the DSA replace or discipline private ordering for trusted flaggers, but instead institute new, parallel structures.

Police Flagging via Internet Referral Units (IRUs)

IRUs are police task forces that perform a flagging role. These flags are called "referrals" since IRUs do not issue legally binding orders; rather, they "refer" content on the ground that it may potentially violate the platform's Terms of Service.[28] As with other forms of flagging, the discretion to moderate remains with the platform. This practice was first initiated by the UK's Counter Terrorism Internet Referral Unit (CTIRU) as early as 2010, and has since been adopted by various governments, including the EU, via Europol.[29] Most of these programs focus on counter-terrorism, though their remit has expanded gradually (for instance, to the

[24] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM (2020) 825 final (Dec. 15, 2020) [hereinafter Digital Services Act].

[25] *Id*.

[26] *Id*. The accompanying recitals to article 19 of the proposal seem to expressly permit IP organizations to gain Trusted Flagger status, with recent amendments adding the caveat that these flaggers must observe "respect for exceptions and limitations to intellectual property rights," which could be comparable to the U.S. doctrine of "fair use."

[27] *Id*., art. 19; *see also* Removals under the Network Enforcement Law, *supra* note 23.

[28] Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 Colum. Hum. Rts. L. Rev. 114 (2018).

[29] *Id*.

combatting of child sexual abuse material, or, in Europol's case, "content promoting illegal immigration services").[30]

Formally, IRUs are not necessarily entitled to priority treatment by platforms. Even without any formal priority rights, however, police flags would be treated with particular care and attention by platforms due to power differentials between police and the average flagger.[31] This has triggered debate about the constitutional and fundamental rights dimensions of voluntary flagging, which we return to further below in 3.2. New frameworks, such as the DSA, foresee IRUs being granted "trusted flagger" status and thus receiving priority treatment as a matter of law. IRUs have also been criticized for a lack of transparency, and it can be difficult to assess the scale and content of their operations.[32] Based on the information available, Europol referred 26,262 items of content in 2019; the Netherlands IRU, by contrast, was far less active, submitting only 1,274 referrals in 2019.[33]

Child sexual abuse material (CSAM)

CSAM has a well-established self-regulatory trusted flagger structure. Its central player is INHOPE, a global network of CSAM hotlines.[34] The National Center for Missing and Exploited Children performs a similar

---

[30] *See* EU Internet Referral Unit, 2020 EU IRU Transparency Report, Europol (Dec. 14, 2021), https://www.europol.europa.eu/publications-events/publications/eu-iru-transparency-report-2020.

[31] *E.g.*, Paddy Leerssen, *Cut Out the Middle Man: The Free Speech Implications of Social Network Blocking and Banning in the EU*, 6 J. INTELL. PROP., INFO. TECH. & ELEC. COM. L. 99 (2015).

[32] *See*, for example, the Israeli case of *Adalah v. Cyberunit*, which illustrates this problem well. Their Supreme Court denied standing because the claimants, public interest litigants, were unable to supply evidence that protected speech had been affected by the IRU's actions. *See* Daphne Keller, When Platforms Do the State's Bidding, Who Is Accountable? Not the Government, Says Israel's Supreme Court, Lawfare (Feb. 7, 2022), https://www.lawfareblog.com/when-platforms-do-states-bidding-who-accountable-not-government-says-israels-supreme-court.

[33] Platform transparency reports offer information on binding government requests, but it is often unclear whether the voluntary 'referrals' issued by IRUs are included in this data. See Government TOS Reports, Twitter, https://web.archive.org/web/20170427171740/https://transparency.twitter.com/en/gov-tos-reports.html (last visited Mar. 13, 2022). The authors were unable to find more recent entries. This section appears to have been subsumed in the general reporting on trusted flagging. Some IRUs also issue their own transparency reports. See EU Internet Referral Unit, *supra* note 30. Additional information has also been gleaned from other public sources such as FOIA requests and parliamentary hearings.

[34] *See* INHOPE, https://www.inhope.org/EN?locale=en (last visited June 27, 2022).

function in the U.S. context.[35] Typically, once content is reported to a hotline, an expert analyst makes an assessment. When they deem the content contains CSAM, a report is made to the police and to the platform, and, simultaneously, the content is hashed and uploaded to databases as reference files to prevent future uploads (this is also known as "notice-and-staydown").[36] By hashing a specific image, an unique ID is created, against which hosting providers and platforms can scan content to automatically remove any content that matches.[37] Further, many tech companies are members of the Technology Coalition, which is a global sector organization aimed at child safety online. Notably, in their annual report, they report that almost 50% of members make use of trusted flaggers.[38]

In this light, trusted flagging practices regarding CSAM appear to be especially sophisticated. One possible explanation for this is that an expert review by a trusted flagger can spare individual platforms' content moderators from having to review this highly traumatic and legally sensitive material. Second, industry, government, and societal interests are aligned in the case of CSAM which greatly incentivizes cooperation. Third, the legal assessment of CSAM is relatively straightforward, as the material is illegal regardless of context and permits no exceptions (in contrast to, for example, hate speech or terrorist content).

Intellectual Property

IP rights-holders are the most influential third-party flaggers. They take part in several sophisticated sector-specific systems.[39] Our analysis here focuses on copyright, though arrangements also exist for other relevant rights, such as trademarks. Available data shows that rights-holders have been extremely active in the use of flagging frameworks such as the DMCA, and the burden created by all these notices may have motivated

---

[35] *See* Nat'l Ctr. for Missing & Exploited Children, https://www.missingkids.org/HOME (last visited June 27, 2022).

[36] *See, e.g.*, Notice and Takedown, INHOPE (Mar. 7, 2020), https://www.inhope.org/EN/articles/notice-and-takedown-ntd.

[37] Robert Gorwa, *Reuben Binns & Christian Katzenbach, Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 Big Data & Soc'y (2020), https://journals.sagepub.com/doi/full/10.1177/2053951719897945.

[38] *See* The Technology Coalition Annual Report, Tech Coalition, https://www.technologycoalition.org/annualreport/ (last visited June 27, 2022).

[39] *See, e.g.*, Multistakeholder Forum on the DMCA Notice and Takedown System, U.S. Patent and Trademark Off., (2015), https://www.uspto.gov/ip-policy/copyright-policy/multistakeholder-forum-dmca-notice-and-takedown-system.

platforms to accommodate them with special privileges.[40] Little is known about these arrangements since they are firmly in the realm of private ordering and have not been guided by government oversight.

However, research confirms that most of the big platforms have created dedicated channels or specific privileges for rights-holders where they can directly notify the platform of content they consider to be infringing on their copyright. Several have even gone as far as to enable rights-holders to directly remove content, automatically complying with their requests and reviewing them only ex post, if at all.[41] The most far-reaching programs have shifted towards proactive filtering based on reference files, comparable to the CSAM strategies described previously. Through its Content ID program, YouTube offers rights-holders the option to automatically detect possibly infringing material upon which the rightsholder can decide to remove, demonetize, or track the content.[42]

Even though the arrangements are all part of the platforms' own policy or self-regulatory initiatives, they are clearly promoted by the strong position legal frameworks afford copyright-holders. Specifically, the statutory threat created by liability regimes such as the US DMCA and the EU Copyright Directive incentivize platforms to treat copyright-holders' notices with priority, relative to other notices. This influential position of rights-holders has been heavily criticized from a free expression perspective: there is hardly any incentive for rights-holders to exercise restraint and invest resources into observing important copyright exceptions and limitations, such as for parodies, pastiche, or citations (or what is known in the U.S. as "fair use"). Finally, the ultimately private relationship between rights-holders and platforms has resulted in very opaque practices. Criticism led YouTube to release its first copyright transparency report in

[40] Daniel Seng, *The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices*, 18 VA. J. L. & TECH. 369 (2014).

[41] Jennifer M. Urban, Joe Karaganis & Brianna Schofield, Notice and Takedown in Everyday Practice, U.C. Berkeley Pub. L. Rsch. Paper No. 2755628 (2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2755628.

[42] *See* Access for All, a Balanced Ecosystem, and Powerful Tools, YouTube (Dec. 6, 2021), https://blog.youtube/news-and-events/access-all-balanced-ecosystem-and-powerful-tools/; How Content ID Works, YOUTUBE, https://support.google.com/youtube/answer/2797370?hl=en#zippy=%2Cwhat-options-are-available-to-copyright-owners%2Crelated-topics%2Cwho-can-use-content-id (last visited Mar. 13, 2022).

2021,[43] which, highlighted how automated systems now outsize manual flagging, and for some authors, confirmed the dangers of over-blocking.[44]

2. Conceptualizing trusted flagging

The discussion above has shown that trusted flagging describes an array of different practices and actors. Building on our working definition of trusted flaggers - i.e. privileged third parties in the flagging process – we are now able unpack this concept along three characteristic dimensions: their legal construction, the nature of their privileges, and the stage of the content moderation process. In other words: their legal status, how they are trusted, and how they flag content.

Firstly, trusted flagging arrangements differ in their legal construction. As mentioned at the start of section 2, all trusted flagger practices are private ordering constructs that can be seen as a logical response to the conditional liability regimes for hosting providers.[45] However, we have also seen that trusted flagging can involve the government to greater or lesser degrees, ranging from co-regulatory to legislative efforts. Government involvement can be seen as particularly significant from a legal perspective, since constitutional free speech norms tend to offer a higher level of protection against restrictions imposed by government than by private actors. Commentators have raised concerns that governments can leverage flagging arrangements to outsource or "privatize" their regulation of speech via private platforms and have argued that government involvement should trigger a higher level of constitutional scrutiny.[46]

The work of Sebastian Schwemer provides a useful framework for thinking through different degrees of public involvement in flagging: on one end of the spectrum are public flagging entities, such as a police IRU.[47] On the other end are strictly self-regulatory private-private frameworks, where private platforms are entering voluntary relationships with other private actors, such as INHOPE. The intermediate model involves explicit public endorsement of a private-private flagging relationship. Our table below expands on Schwemer's model by distinguishing between

[43] *See Access for All, a Balanced Ecosystem, and Powerful Tools*, *supra* note 42.

[44] Paul Keller, *YouTube Copyright Transparency Report: Overblocking Is Real*, KLUWER COPYRIGHT BLOG (Dec. 9, 2021), http://copyrightblog.kluweriplaw.com/2021/12/09/youtube-copyright-transparency-report-overblocking-is-real/.

[45] *See About the YouTube Trusted Flagger Program*, *supra* note 8.

[46] Schwemer, *supra* note 3.

[47] Schwemer, *supra* note 3.

flagging based on unlawful content which can trigger liability, and flagging based on the terms of service, so flagging possible lawful content.[48] This is an important distinction because the threat of liability can be an important motivator for platforms to heed third-party demands and accommodate them with trusted flagging privileges. In this sense, only private flagging without a threat of liability is entirely voluntary.

Model	**Terms of Service**	**Liability**
Private flagger	Hate Speech	Copyright holders; INHOPE
Private flagger with public endorsement	'Trusted flaggers' appointed under the Digital Services Act or NetzDGauto	
Public flagger	Police IRU	

Fig. 1 – Model for government involvement in Trusted flagger practices (adapted from Schwemer 2018).

Second, trusted flagging differs significantly in terms of its flagging mechanisms and their position in the overall content moderation process. Clearly, one can distinguish trusted flagging from third-party involvement at earlier stages of the content moderation process. For instance, third parties can also be involved ex post in arbitrating appeals, or ex ante as advisors for policy drafting and standard setting.[49] However, other types of third-party involvement are less easily distinguished, and start to overlap with trusted flagging. For instance, Facebook's fact-checking partners assess content for potential action by Facebook, but they differ from trusted

[48] The precise conditions for liability differ between jurisdictions, depending on their intermediary liability frameworks. In the EU and many other jurisdictions worldwide, platforms can become liable for all types of illegal content once they obtain actual knowledge of its presence on their service. In this context, even flagging based on Terms of Service has the potential to trigger liability, if it brings unlawful content to the platforms' attention. Under U.S. law, by contrast, flagging based on Terms is more clearly separate from legally binding notices. Under Section 230, platforms are immune as regards user content. The only major exceptions to this regime are for copyright and trademark claims (these notices must also adhere to the specific takedown notice format of the Digital Millennium Copyright Act), as well as for federal criminal law.

[49] For a discussion of civil society's diverse engagements in the "networked governance" of platforms, see Robyn Caplan's contribution to this Essay Series. [reference t.b.a.]

flaggers in that Facebook supplies fact checkers with a feed of potential items to review.[50] Here the content has already been detected and classified in some preliminary way by the platform, and the fact-checker takes up a hybrid position somewhere between an external notifier and an external moderator. Furthermore, trusted flagging entities can also simultaneously perform other roles in platform governance. For instance, the same NGOs that flags content may also provide input on more high-level policymaking. There are also significant overlaps with policework and surveillance. Similarly, NGOs in areas such as hate speech may also have policies or may be required to refer cases to law enforcement.[51]

Third, we have seen that different flagging arrangements involve widely different privileges, coordinating with platforms more or less closely. Trusted flaggers receive some degree of priority or expedited review, but how this works out in practice may differ. It is also conceivable that trusted flaggers are subjected to lower substantive standards of review, although we have encountered formalized policies to this effect. Flaggers can also receive other privileges besides priority review. For instance, some flaggers can access specific interfaces or communication channels that help them submit notices at scale. These arrangements are often buttressed by (formal and informal) feedback and engagement from the platform, which is rarely available to the average flagger.

In the most extreme cases, third parties are granted such extensive privileges that they start to look less like conventional flagging and more like a wholesale outsourcing of content moderation decisions. One study suggests that flags from certain copyright holders are not reviewed at all, being granted automatic deference, and only being reviewed ex post or perhaps not at all.[52] In this instance, the party acts both as flagger and as moderator at once. Even further removed are automated hashing databases, where third parties can submit reference files to the platform for purposes of automated content filtering. Here, the third party can effectuate content removals without even needing to reference or 'flag' any specific content on the platform. These more advanced arrangements speak

[50] Mike Ananny, *Checking in with the Facebook Fact-checking Partnership*, Colum. Journalism Rev. (Apr. 4, 2018), https://www.cjr.org/tow_center/facebook-fact-checking-partnerships.php.

[51] Recent proposals, such as the revised NetzDG and the proposed DSA, envisage expanded duties to refer unlawful content to police. *See* Digital Services Act, art. 19, *supra* note 19.

[52] Such as the U.S. IP flagging arrangements known as "DMCA Auto." Jennifer M. Urban, Joe Karaganis & Brianna L. Schofield, *Takedown in Two Worlds: An Empirical Analysis*, 64 J. COPYRIGHT SOC'Y 483 (2018).

to something of a paradox in the context of trusted flagging: the more a party is trusted, the less it needs to flag.

3. Discussion: Three narratives about trusted flagging

Trusted flaggers elicit competing narratives about their role in platform governance. Their proponents typically defend trusted flaggers as a means to outsource knowledge or decentralize control in content moderation. In response, a more critical counternarrative has highlighted how trusted flagging arrangements reflect and reinforce pre-existing power structures, including state coercion and private power. This discussion revolves around two competing views of trusted flagging: either as trustworthy experts working to make content moderation more effective and legitimate, or as self-interested co-opters spurring its worst excesses. We propose an additional, third perspective, which views trusted flagging partnerships as essentially performative.

3.1 The trusted flagger as a source of expertise and inclusion

Sebastian Schwemer has observed that trusted flaggers' legitimacy rests on claims of representativeness, whether it be of the democratic state, a private right holder, or an NGO with knowledge of specific communities, cultures, or interest groups. On this basis, they can appeal to a particular expertise or normative standing in assessing online harms. This line of reasoning responds to the criticisms that platforms have centralized too much control over online speech. Furthermore, these platforms commonly fail to incorporate (local) contexts and cultures, as may often be required on issues such as hate speech or disinformation.[53] It also resonates with the ideal of multi-stakeholderism, which has historically been central to internet governance and is now, as Robert Gorwa observes, a driving force behind platforms' engagement with NGOs.[54]

---

[53] Schwemer, *supra* note 3.

[54] Robert Gorwa, *The Platform Governance Triangle: Conceptualizing the Informal Regulation of Online Content*, 8 Internet Pol'y Rev. (2019), https://policyreview.info/node/1407.

Marginalized groups in particular face difficulties in obtaining adequate protection and redress from platforms.[55] For example, platforms' "race-blind" content moderation policies and its policies that center specific content rather than patterns of abuse render many harms invisible or incontestable.[56] In theory, the trusted flagger model could allow representative interest groups to acquire a stake in platform moderation and give voice to voiceless parties. In this way, trusted flaggers could potentially help decentralize the control platforms have over online speech and even out existing power structures. However, policymaking thus far has not engaged in much depth with eligibility or capacity criteria for trusted flagging.[57]

Representation typically refers to large groups, such as democratic polities, racial or ethnic groups, LGBTQIA+ groups, and so forth. But it should be noted that flaggers can also directly represent individuals. The most common example from practice is IP rights-holders, but one might also envisage trusted flaggers supporting individual victims of online harms, such as victims of online harassment, abuse, non-consensual sexual imagery and so forth. Extensive scholarship on access to justice has shown that often people lack the wherewithal to exercise their rights, and locally embedded flagging entities could thus play an important facilitative role.[58]

3.2 The trusted flagger as unaccountable co-optation by public and private power

---

[55] *See, e.g.*, Caitlin Ring Carlson & Hayley Rousselle, *Report and Repeat: Investigating Facebook's Hate Speech Removal Process*, 25 First Monday (2020), https://journals.uic.edu/ojs/index.php/fm/article/view/10288; Bharat Ganesh, *Platform Racism: How Minimizing Racism Privileges Far Right Extremism,* Items (Mar. 16, 2021), https://items.ssrc.org/extremism-online/platform-racism-how-minimizing-racism-privileges-far-right-extremism/; Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children*, ProPublica (June 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

[56] *See, e.g.*, Ángel Díaz & Laura Hecht-Felella, *Double Standards in Social Media Content Moderation*, Brennan Ctr. for Just. (Aug. 4, 2021), https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation.

[57] The DSA's approach can be described as rather technocratic in that it emphasizes "expertise" without reference to any other forms of (cultural, political, socioeconomic) representativeness.

[58] Naomi Appelman et al., *Access to Digital Justice: In Search of an Effective Remedy for Removing Unlawful Content*, in Frontiers in Civil Justice: Privatization, Monetisation, and Digitisation (X. Kramer et al. eds., forthcoming 2022).

More critical perspectives on trusted flagging warn against over-blocking and a lack of accountability. For many commentators they aggravate, rather than assuage, concerns about the lack of transparency, accountability, and contestability of content moderation practices.[59]

This critique can be articulated from a governance perspective and from a political-economic perspective. From a governance perspective, Schwemer argues that trusted flagging entails a risk of institutional bias against freedom of expression, since trusted flaggers have a mandate to effectuate removal but no requirement to protect freedom of expression. More generally, Brenda Dvoskin has shown how advocacy groups pushing for more restrictive and "aggressive" content moderation have been more successful than those aiming to protect the freedom of expression.[60] Institutionalized trusted flaggers programs could, if anything, exacerbate these trends. This misalignment of incentives might be especially problematic when an actor's role as trusted flagger aligns with its economic interests (e.g., IP rights-holders) or bypasses constitutional safeguards for government action (i.e., 'privatized censorship').[61] Platforms may lack adequate incentives to combat such over-removal, since there are typically no legal constraints on their power to remove content. Indeed, the platform might try to justify its actions by placing responsibility for their removal actions with the trusted flagger. Finally, concerns have also been raised about the lack of transparency and accountability in trusted flagging arrangements: at a systemic level, essential information is often lacking, such as the identity of the flaggers involved and the extent of their activity. At the individual level, users who are engaged in flagging are usually not notified when the platforms respond to the flagged content. Furthermore, since the involvement of trusted flaggers remains opaque, users have no ability, legal or otherwise, to contest trusted flagging or hold it accountable.[62]

A political-economic perspective highlights how trusted flagging reflects, and works in service of, pre-existing power structures. The ideal of representative civil society groups volunteering to take on content

---

[59] *See* Schwemer, *supra* note 3; Federica Casarosa, When the Algorithm Is Not Fully Reliable, in Constitutional Challenges in the Algorithmic Society 298 (Hans-W. Micklitz et al. eds., 2021).

[60] Brenda Dvoskin, *Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, VILL. L. REV. (forthcoming 2022).

[61] Martin Husovec, *Accountable, Not Liable: Injunctions Against Intermediaries*, TILEC Discussion Paper No. 2016-012 (2016), https://papers.ssrn.com/abstract _ id=2773768.

[62] *See* Seng, *supra* note 40.

moderation duties belies the fact that there is no obvious funding model for such activity. Our review of empirical evidence shows that trusted flagging works primarily in service of vested public and private powers, in particular law enforcement and IP rights-holders. These parties can muster the political clout to demand recognition from platforms and the wherewithal to engage in large-scale monitoring and reporting in line with their political and economic interests. In some cases, trusted flaggers might also be compensated by the platform itself, raising questions about their independence. Also, engagement with civil society groups often requires encouragement from governments, as in the discussed EU Code of Conduct on speech, and even then, there is little evidence that these arrangements have at all approached the scale and influence of flagging by police, and, especially, IP-holders.

These perspectives have prompted debate about the need for safeguards in trusted flagging, such as eligibility criteria for trusted flaggers, performance reviews and oversight, and both public and individual transparency.[63] As indicated, the proposed DSA does try to establish several of these safeguards but, crucially, only does so for the trusted flaggers appointed by the EC. In other words, the DSA institutes a new, parallel structure for trusted flagging, but it does not seek to regulate existing arrangements.[64] Still underexamined, therefore, are possibilities to regulate existing trusted flaggers and introduce new safeguards, for instance by setting conditions for their transparency (e.g., should NDAs be permitted? do affected users deserve to be notified about trusted flags?) and accountability (e.g., Who should fund flagging? Can flaggers be liable for errors? Might users appeal their decisions?). An overarching question is whether trusted flaggers might be regulated directly, by way of their own legal duties and oversight structures, or only indirectly, by imposing duties on platforms to introduce appropriate safeguards.

3.3 The trusted flagger as performance of inclusion

[63] *See* Husovec, *supra* note 63; Schwemer, *supra* note 3; *Communication from the Commission*, *supra* note 4.

[64] Article 20 of the DSA does provide that platforms have to take appropriate measures to prevent the misuse of their reporting mechanisms. *See* João Pedro Quintais & Sebastian Felix Schwemer, *The Interplay Between the Digital Services Act and Sector Regulation: How Special Is Copyright?*, EUR. J. OF RISK REGUL. (forthcoming 2022). However, these only apply in cases of manifest misuse.

A connected line of critique suggests that trusted flagging is not so much illegitimate or unaccountable as it is insignificant. Automation looms over all debates around flagging. For example, in the first half of 2020, 96.4% of the content removed by TikTok was found by its own automated systems before any user reported it.[65] Similarly, of all the hate speech Facebook removed in 2021, 96.5% was found by the platform itself.[66] Even though the platforms do not publish specific numbers on the amount of notifications received by trusted flaggers, we can gather from the general numbers alone that their impact in terms of volume on the overall moderation process cannot be large.[67]

In short, trusted flagging does not scale. If third parties wish to influence content moderation as it is currently practiced, they must leverage its automation. If flagging is to play a significant role going forward, it will, at a minimum, be through require notice-and-staydown approaches, where action is taken not only against the flagged item but also to equivalent and future uploads of the same material. Trusted flagging programs that are embedded in this automated content moderation process have a much higher chance of systematically influencing content moderation.[68] As discussed, these most influential arrangements are not really 'flagging' specific positions at all, but feed reference files directly into the platform's automated removal logics. These developments cast trusted flagging itself in entirely different light: a tinkering around the edges of content moderation rather than a true shift in power relations.

How then might we explain the rise of trusted flagging, and all the attention it garners? The motivation for platforms to engage in trusted flagger partnerships may lie primarily in its symbolic value.[69] Trusted flagging can allow platforms to perform multi-stakeholderism, inclusion, and reform, whilst leaving the core of their operations untouched. In short, it is a PR move which offers platforms legitimacy but does not substantially alter their content moderation practices. Indeed, in their public communications it appears that platforms are relatively forthcoming and open about their "partnerships" with NGOs, and far less so about their partnerships with police and IP right-holders. Opacity about the relative lack of

---

[65] *See* Community Guidelines Enforcement Report, *supra* note 12.

[66] *See* Hate Speech, Meta https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/ (last visited Mar. 13, 2022).

[67] *See* Reynders, *supra* note 19.

[68] Such as CSAM and copyright infringing material that can be removed via, respectively, Photo DNA or Content ID.

[69] Dvoskin, *supra* note 62.

impact from NGO flagging can help platforms to profit from the legitimacy or goodwill associated with these connections while not giving up any meaningful control over their content moderation practices.[70] To the extent that these third parties can influence moderation at all, it is more likely to be through other avenues, such as high-level policy consulting.

Conclusion

This essay has shown how the concept of "trusted flagging," in its everyday operationalization, serves as a site of contestation between competing interests and legitimacy claims in platform governance. This single label covers a great diversity of third-party flagging constructions, in service of many different interests from law enforcement to NGOs to IP holders. We also see great disparities: legal constructions and automated systems play a crucial role in shaping the influence and effectiveness of different flaggers. The most influential strategies leverage automation: they ensure that their flagging actions are scaled up through stay-down mechanisms or have foregone the flagging modality entirely in favor of automated filtering based on reference files. Conventional flagging by third parties, to put it bluntly, does not scale, and there is little hope that it will fulfil its promise of more decentralized, legitimate, inclusive content moderation. To the extent that it has influenced platforms, it has done so primarily in service of existing power structures, such as the state and IP industry, or by entrenching the position of the platforms themselves by increasing the perceived legitimacy of their content moderation practices.

We see several ways forward. Regulation has a clear role to play, but we propose that attention should shift away from creating yet more parallel flagging structures—as the NetzDG already has and the Digital Services Act now foresees—towards ensuring greater scrutiny of existing structures in regulation and private ordering. As platforms further accommodate, integrate, and automate the demands of powerful third parties such as law enforcement and the IP industry, additional safeguards are essential to prevent overreach. (At a bare minimum, for instance, removal decisions instigated by trusted flaggers should be notified to affected users, so that they can contest these decisions at their source. Non-disclosure agreements for trusted flaggers are another point of grave concern.) In theory, the trusted flagger model may still hold promise as a source of inclusion in platform governance; trusted flaggers may act as points of

[70] Sarah T. Roberts, *Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation*, 23 First Monday (2018), https://www.firstmonday.org/ojs/index.php/fm/article/view/8283.

contact or helplines for victims of online harms and help them navigate platform moderation procedures effectively. But to become reality, such an approach would require far more sustained investment in the capacity and visibility of flagging organizations. In any case, it should not be mistaken for a true decentralization of power in platform governance, which now occurs, if at all, at the level of automated content filtering.