4. Constitutional Aspects of Trusted Flaggers in the Netherlands

Jacob van de Kerkhof

Abstract: The use of trusted flaggers is an established practice in content moderation by internet intermediaries such as Meta and Google. It allows engagement with expertise of governmental and non-governmental organizations, ensuring swift actionability of flagged content. The Digital Services Act formalizes this practice in Article 22. State entities have also been functioning as trusted flaggers, which has been a topic of scholarly and societal debate. This chapter discusses the constitutional tensions of the existing and new Digital Services Act (DSA) framework of trusted flaggers in the Netherlands with the right to freedom of expression as laid down in Article 7 of the Dutch constitution and Article 10 of the European Convention on Human Rights (ECHR). It makes several suggestions to increase the lawfulness, legitimacy, and accountability of this framework.

Keywords: content moderation, flagging, freedom of expression, Digital Services Act (DSA), accountability, transparency

Introduction

Over the past decades, the public debate has moved to the digital realm, in part to social media (Balkin 2018). Social media are governed by internet intermediaries such as Meta and Google, who are commercially motivated private entities. Social media spaces have greatly expanded the possibilities for freedom of expression, due to the increased reach that anyone can achieve. Over time, the risks of social media have also become apparent: the anonymous sharing of illegal and harmful content has real-world effects (Arcila and Griffin 2023). These risks call for a more public values–based

Van Dijck, José, Karin van Es, Anne Helmond, and Fernando van der Vlist, eds. *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values*. Amsterdam: Amsterdam University Press, 2025
DOI 10.5117/9789048562718_CH04

approach to social media governance, creating a tension with the commercial nature of internet intermediaries governing social media. The process of governing speech in social media spaces is called content moderation; it involves norm-setting and enforcement thereof by internet intermediaries of third-party-generated content and ordering its display to the public (Gillespie 2018). Internet intermediaries have sought ways to legitimize their content moderation processes to create a more public values–based approach to combat the risks associated with harmful and illegal content online. One of these ways is involving external parties, in a paradigm that Caplan (2023) likens to networked governance. Networked governance is a term coined to capture the paradigm of leveraging various actors beyond traditional governmental bodies in governing societal issues, thus decentralizing the power away from a single authority, in this case the social media platform. In content moderation, external actors are involved both in norm-setting, for example, by involving civil society organizations to express concerns for minority interests in community guidelines, as well as in enforcement, for example, by providing fact-checking.

Internet intermediaries have engaged external organizations in the detection of illegal and harmful content for quite some time, a phenomenon referred to as "trusted flaggers" (Eghbariah and Metwally 2021). The term is derived from the concept of "flagging," a mechanism for reporting offensive content to a social media platform through expressing concerns within the predetermined rubric of a platform's community guidelines (Crawford and Gillespie 2016). Anyone can flag content, which allows platform users to engage in content moderation. This democratizes the content moderation process, but abuse to deplatform other platform users has been reported, for example (Are and Briggs 2023). Trusted flaggers are organizations with expertise in a particular content area that are granted priority access ("trusted") to "flag" illegal content (as opposed to flags from "ordinary" users). The internet intermediary expeditiously reviews trusted flags and determines whether content remains accessible, is taken down, or faces another form of sanction. This process has the advantage of legitimizing the internet intermediaries' content moderation process, because of the expertise of flagging organizations and potential for representation of minority interests through trusted flagger organizations (Appelman and Leerssen 2022). Treating "trusted flags" more expeditiously means that illegal content can be removed more quickly, making the social media platform a safer—and therefore more attractive—venue for users and, crucially, advertisers (Griffin 2023). Overall, trusted flaggers are seen as a positive exponent of networked governance, which is underlined by the

formalization of the trusted flagger system in Article 22 (see also rec. 61–62) of the Digital Services Act (Regulation 2022/2065, DSA), the European Union's latest instrument to regulate internet intermediaries. Currently, trusted flagger arrangements are voluntary. Following the DSA, the newly appointed national digital services coordinator (DSC) can appoint organizations as "trusted flaggers," and social media platforms need to accommodate this.

This chapter focuses on state entities functioning as trusted flaggers. In principle, trusted flaggers can be any type of entity. The most common examples are NGOs, national police, and state bodies. Especially in the case of the latter two, this creates tension, since state entities have to respect fundamental rights in interacting with citizen's speech, which includes content on social media platforms. Although internet intermediaries make the final call on third-party-generated content, a referral by a state entity can be conceptualized as a strong nudge to remove that content (Bambauer 2015; Kreimer 2006). In fact, Urban et al. (2017) found that flagging by trusted flaggers can lead to removal without review in some cases. This raises a fundamental rights concern: if a state actor requests removal of third-party-generated content, and if this request pressures the internet intermediary to remove that content, sometimes even without review, that state actor might be engaging in what Kreimer describes as "censorship by proxy" (2006). Crucially, this creates tension with the freedom of expression as laid down in Article 7 of the Dutch constitution and Article 10 of the European Convention on Human Rights (ECHR). The purpose of this chapter is to evaluate the freedom of expression in the Dutch constitutional setting in the context of the trusted flagger framework.

The contribution starts with a description of standing trusted flagger practices in the Netherlands as well as a description of that framework in the DSA. Next, it describes the embedding of freedom of expression in the Dutch constitution and the ECHR. Subsequently, it synthesizes these sections, assessing whether there are any constitutional fragilities to trusted flaggers in the content moderation process. Finally, it makes several suggestions to increase the lawfulness, legitimacy, and accountability of this framework.

Trusted flaggers: An introduction

Internet intermediaries occupy a crucial role in moderating the public debate on social media, placing them in a quasi-public position that requires them to take responsibilities usually reserved for states (Klonick 2018). In taking this responsibility, internet intermediaries have sought to legitimize content

moderation by seeking external validation. External adjudicatory bodies such as the Oversight Board, engagement with fact-checking organizations, or X's Trust & Security Council (dissolved at the end 2022) are examples of such external validation. Trusted flaggers also fall within this concept. For the purpose of this contribution, a trusted flagger is defined as any entity that flags content through privileged channels for the internet intermediary to review. Trusted flaggers can be private, semi-private, and public bodies that have been enlisted in a privileged flagging capacity based on their societal interest, legal interest, or expertise. Appelman and Leerssen (2022) identify three distinguishing characteristics: (1) the legal status of the trusted flagger; (2) the stage of involvement in the content moderation process; and (3) the degree of privilege in their flagging practice. This section discusses the legal status of flaggers and the degree of their privileges. It excludes the stage of the content moderation process, as this contribution solely focuses on flagging after the content is published. Trusted flagger arrangements vary widely. In some instances, trusted flaggers may be involved on a bilateral voluntary basis. For example, YouTube has an outreach program by which it allows organizations with certain expertise to aid via prioritized flagging tools. In other instances, cooperation is semi-voluntary. Internet intermediaries have opted to join co-regulatory instruments that create a role for trusted flaggers, such as the EU's "Code of Conduct on Countering Illegal Hate Speech Online"—in which they are referred to as "trusted reporters" (EU Code of Conduct 2016, 3)—and the Strengthened Code of Practice against Disinformation (Commitment 21). Those instruments encourage internet intermediaries to create a position for trusted flaggers in their content moderation process. The same goes for self-regulatory instruments, for example, the Technology Coalition against child sexual abuse material.

Sometimes the trusted flagger has a particular right to enforce against content online, for example, in cases of intellectual property enforcement. Copyright protection organizations and individual rights holders function as trusted flaggers. YouTube offers ContentID for copyright holders but also reports direct relations with rightsholders. The police require special attention as trusted flaggers within the content moderation process. Under the DSA, law enforcement can engage in two different interactions: firstly, it can issue takedown orders for specific content based on national or EU law under Article 9 of the DSA through the DSC. In those instances, social media platforms are legally obligated to comply with the takedown order. Secondly, the police can also serve as trusted flaggers by referring content to internet intermediaries for review. These police bodies have been dubbed "internet referral units" and can be seen, for example, in the United

Kingdom, Europol, and Israel (Chang 2017). Finally, next to these various legal statuses of trusted flaggers, there have been efforts to legalize the position of trusted flagger in national law. For example, in the case of the German Netzwerkdurchsetzungsgesetz (Network Enforcement Act), the German legislature formalized the option to flag on the grounds of public interest, functioning as a reporting agency (*Beschwerdestelle*). Public interest flags are subject to transparency requirements, with internet intermediaries having to disclose how many public interest flags they receive. The latest formalization in the Digital Services Act is discussed in the next subsection.

The second differentiating factor is the stage at which the trusted flagger is involved. While the name suggests that they are only involved in "flagging" content, meaning that after the content is posted, trusted flagger organizations can also be involved in policymaking, representing specific interests in creating community guidelines. The involvement of civil society organizations in forming community guidelines is encouraged under Article 46 of the DSA. Since this is not specific to the trusted flagger functions discussed in Article 22 of the DSA, this stage is not treated in this contribution.

The third feature differentiating trusted flaggers is their degree of privilege with the internet intermediary. Trusted flaggers have different levels of access to the internet intermediary, which is also dependent on their legal status. These range from treating the flag almost as a standard content flag, with little urgency or lessened discretion for the platform, to situations where the review of a flag from a trusted flagger is reduced to a bare minimum, as seen with copyright holders under the US Digital Millennium Copyright Act (Urban et al. 2017). The difference in privilege depends on the expertise of the trusted flaggers and the potential consequence of disregarding the referral: as mentioned earlier (Kreimer 2006), a critique of state bodies referring content to internet intermediaries for review is that such a referral exudes pressure for internet intermediaries to remove that content, which is difficult to resist. Bambauer coins this phenomenon "jawboning"—encapsulating both formal and informal pressure to comply with a state's bidding (2015).

Trusted flaggers in the Digital Services Act

The DSA formalizes the trusted flagger system. Trusted flaggers are appointed by the DSC based on their expertise, independence, and diligence (Article 22(2)). The DSC must disclose trusted flaggers it appointed to the European Commission, and this information is made public. Additionally, the process of flagging has also been formalized in Article 16, which pertains to notice and action mechanisms. Article 16 mandates internet intermediaries

to allow all entities, such as users, interested parties, and government officials, to flag content they deem illegal. The illegality of content must be based on potential violation of EU law or national law in accordance with EU law (Article 3(h)). The difference between flags *ex* Article 16 and trusted flags *ex* Article 22 is the requirement that trusted flags are treated without undue delay, whereas flags *ex* Article 16 need to be treated in a timely, diligent, non-arbitrary, and objective manner (Article 16(6)). Further, trusted flaggers must publish a public report of the notices they have filed every year and send that report to the DSC (Article 22(3)). Under certain conditions, trusted flaggers may be stripped of their status when they are no longer deemed to fulfill their function well (Article 22(7)).

Essentially, the Digital Services Act codifies and formalizes a standing practice. This formalization is noteworthy for several reasons. Firstly, appointing trusted flaggers so far has been a voluntary arrangement, happening exclusively in the sphere of private law. When the DSC—which is a state entity, e.g., in the Netherlands, the Authority for Consumers and Markets (ACM)—appoints trusted flaggers, the arrangement with the internet intermediary becomes compulsory. This raises questions regarding the public law responsibilities and accountability of the DSC, including the actionability of the decision to appoint trusted flaggers, or not to grant that status. Secondly, the formalization has a practical concern: the Digital Services Act does not preclude internet intermediaries from maintaining existing trusted flagger relations; it only ensures that the DSC has the capacity to add to those arrangements (DSA rec. 61): "In particular, industry associations representing their members' interests are encouraged to apply for the status of trusted flaggers, *without prejudice to the right of private entities or individuals to enter into bilateral agreements with the providers of online platforms*." Although this means that there is increased transparency on the to-be-appointed flaggers, it does not diminish the opacity of current arrangements, adding an extra layer to the abovementioned networked governance. The question is whether appointment through the DSC—although compulsory for internet intermediaries—can serve as an appealing avenue for entities seeking to be trusted flaggers. In current arrangements, those entities can flag content based on national law and community guidelines, whereas the trusted flagger framework proposed in the DSA only allows for flagging of illegal content *ex* Article 3(h) covering only content in violation of national or EU law. This means that trusted flaggers under the DSA may only flag a limited scope of content—only that which violates national or EU law, not that which violates terms and conditions. It is expected that in practice, this distinction does not lead to limitations, but formally, DSC-appointed

trusted flaggers are afforded less possibilities than trusted flaggers through existing arrangements.

Trusted flaggers in the Netherlands

For this chapter, it is important to differentiate between governmental and non-governmental entities functioning as trusted flaggers. Both function as trusted flaggers, yet for the constitutional angle of this contribution, the focus is on governmental organizations: constitutional and fundamental rights norms do not necessarily apply to non-governmental entities.

In the Netherlands, several members of parliament (MPs) have requested transparency on the role of Dutch governmental bodies as trusted flaggers. In 2023, Minister of the Interior and Kingdom Relations Hanke Bruins Slot disclosed which organs of the Dutch government had access to Meta's trusted flagger portal (Kamerstukken 2022–23, no. 1599). The Dutch Ministry for Internal Affairs was the prime addressee of those questions, considering its quest for combating disinformation. Despite the sensitivity of the topic of the requests, the ministry receives or reports a relatively low volume of notifications. In December 2022, Minister Slot reported four cases since acquiring trusted flagger status for Meta-associated platforms in 2019 and two cases to Twitter. Most cases dealt predominantly with disinformation around elections, which falls under the jurisdiction of the Ministry of the Interior and Kingdom Relations. The content identified by these flags concern voting procedures, for example, suggesting that casting a vote would give permission for vaccination. Excerpts from the content removal requests show that internet intermediaries rejected the government's requests, challenging the hypothesis that referrals from state bodies exert pressure on the internet intermediaries to remove content. Meta refused removal because the ministry's interpretation of community guidelines differed from its own.

Aside from the Ministry of the Interior and Kingdom Relations, the national police also received attention in its role as trusted flagger (*Aanhangsel Handelingen* II 2022–23, no. 1946). Questions regarding their role in the content moderation process, raised by conservative MP Pepijn van Houwelingen, primarily focus on the relative opacity of their content removal requests. The police do not keep track of their removal requests, nor does Dutch law require internet intermediaries to do so. As a result, it is unclear what content the removal requests are based on. According to the literature, police units have expressed interest in tackling terrorist propaganda and child sexual abuse material (Kilpatrick and Jones 2022).

Oversight bodies such as the Food and Consumer Product Safety Authority (ACM), the Gambling Authority, and the Authority for Financial

Markets (AFM) comparatively flag a lot more content than the Ministry of the Interior and Kingdom Relations: the gambling authority has flagged seventy-three pieces of content since 2020, the AFM flagged 134 pieces of content in 2019 alone. The Gambling Authority targets illegal forms of gambling, predominantly fake lotteries. It does so by using Meta's Gambling Regulatory Channel, a priority access portal designed for gambling authorities, but it refuses to disclose the exact process of its flagging. The Gambling Authority bases its authority on Article 33b of the Wet op de Kansspelen (Gambling Act). The AFM flagged content related to fake or malicious financial products, requesting removal of 134 pieces of content in 2019. It has since stopped using its trusted flagger status, since the platform's search algorithm has since made it more difficult to track pieces of illegal content. The AFM bases its trusted flagger activities on enforcing the Wet op het Financieel Toezicht (Financial Supervision Act). The ACM is tasked with acting against harmful products and misleading advertisements. It bases its enforcing powers on EU Regulation 2019/1010 on product compliance. Although the ACM did not track the number of requests it made as a trusted flagger, the increasing commercialization of social media spaces (Goanta 2023) raises the suspicion that the amount of potential flags is large. In a landscape in which goods are increasingly being sold on the internet, and consumers are increasingly involved in selling those goods (Mak 2022), it is expected that the consumer authority needs to exercise all available oversight capabilities (Goanta and Spanakis 2022).

As for non-governmental bodies acting as trusted flaggers, it is difficult to create a full list of Dutch non-governmental entities with a trusted flagger position. As mentioned earlier, social media platforms are secretive about who has access to priority notice-and-takedown avenues. NGOs do not always advertise their position as trusted flaggers either. Some Dutch organizations have identified themselves as trusted flaggers, such as PersVeilig (PressSafe) and the Expertisebureau Online Misbruik (Expertise Agency Online Abuse), which focus on issues related to online safety and abuse. Most of those organizations have strong relations with governmental bodies but can still be considered NGOs.

The protection of freedom of expression in the Netherlands

This section introduces the right to freedom of expression in the Netherlands, to offer background to the fragilities to this right in the trusted flagger framework explored in the next section. This right is predominantly

safeguarded through two documents: the Dutch constitution (Grondwet voor het Koninkrijk der Nederlanden [Constitution of the Kingdom of the Netherlands] or Gw) and the European Convention on Human Rights (ECHR).

Article 7 of the Gw safeguards freedom of expression and consists of four provisions. The initial three provisions affirm that individuals do not need prior permission to expose, publish, or broadcast their thoughts or opinions through different media types (Hins 1995). Expressions on the internet are covered by the third sub, adding the exception that each person must act without prejudice to their responsibility under the law. The phrasing of the article is peculiar: the right to freedom of expression is such that one does not need to ask permission to express oneself. The right to freedom of expression covers a right to express, but also a right to disseminate those expressions (De Meij et al. 2000; see also Hoge Raad, November 7, 1892, *Haagse Ventverordening*). The right to disseminate one's expression can be limited by law, but there must always be a meaningful alternative available to spread one's expressions (Hoge Raad, April 26, 1996, *Rasti Rostelli*; see also the European Court of Human Rights [ECtHR], May 6, 2003, *Appleby v. United Kingdom*).

The primary limitative ground of the right to freedom of expression *ex* Article 7(3) of the Gw is everyone's responsibility under the law. Article 7 of the Gw protects shocking and hurtful expressions, provided they add to the public debate (Hoge Raad, January 9, 2001, *van Dijke*). This notion is based in the ratio that freedom of expression is necessary for a functioning democracy; expressions devoid of meaning, such as throwing paint bombs (Hoge Raad, April 19, 2005, *Verfbom*) or sending spam messages (Hoge Raad, March 12, 2004, *Xs4All*).

Article 10 of the ECHR protects the right to freedom of expression on a European level. Because Article 7 of the Gw is not directly enforceable in Dutch courts due to the prohibition on constitutional review *ex* Article 120 of the Gw, most case law in the Netherlands on freedom of expression is based on the ECHR. Article 10 of the ECHR has two parts: Sub 1 provides everyone with the right to freedom of expression, to hold opinions and to impart information and ideas without interference. Sub 2 provides the limitative grounds to that right: the right can be subject to restriction, if such restriction is prescribed by law, serves a legitimate aim, and is necessary in a democratic society. This also applies to expressions on the internet, such as the use of platform affordances (ECtHR, September 15, 2015, *Melike v. Turkey*) and content moderation policies (ECtHR, June 16, 2015, *Delfi v. Estonia*). The provision of Article 10 of the ECHR has a wide scope and covers expressions that may "offend, shock, or disturb the State or any sector of the population"

(ECtHR, December 7, 1976, *Handyside*). The protection of Article 10 of the ECHR also encompasses the right to receive information: for example, in *Yilderim v. Turkey* the ECtHR found that blocking access to a social media platform violates the right to freedom of expression. In that case, disabling Google did not allow citizens to be informed as to effectively exercise a right to freedom of expression. Article 10(2) provides reasons for which the right to freedom of expression may be interfered: interference must be provided for by law, be necessary in a democratic society, and pursue one of the legitimate aims listed exhaustively in Article 10(2) of the ECHR. These tests ensure that an interference is legally foreseeable, proportional, and suitable to achieve its societal goal (ECtHR, April 22, 2013, *Animal Defenders International v. The United Kingdom*).

The fragility of the right to freedom of expression in the trusted flagger framework

A referral by a trusted flagger might impair an internet user's freedom of expression. It is the internet intermediary who has the most profound impact on the freedom of expression of internet users: it has the final say on whether content is accessible or not. Since social media platforms are private entities, they do not need the same regard to a user's freedom of expression: fundamental rights do not apply to internet intermediaries as they do to states (Teubner 2017). Therefore, the freedom of expression does not pose constitutional concerns when social media platforms engage in content moderation. However, the act of flagging a piece of content by a state entity can result in what Kreimer (2006) calls "censorship by proxy": the internet intermediary succumbs to the pressure of the trusted flagger to remove content. Pressure emitting from such a nudge might be difficult to resist (Bambauer 2015), causing freedom of expression concerns. Kaye (2019) reports that internet intermediaries have yielded to government pressure from totalitarian states to silence minority voices. The indirect pressure emitted from a state body acting as a trusted flagger might violate the right to freedom of expression. This fragility is explored in light of the limitation grounds of Article 7 of the Gw, namely lawfulness, and Article 10 of the ECHR, legality, necessity in a democratic society and legitimate aim, respectively.

To create an overview of potential fragilities, one can derive four scenarios from the description above: (1) state actors functioning as trusted flaggers under Article 22 of the DSA; (2) a state actor functioning as trusted flagger outside of the DSA, in a private agreement with the internet intermediary;

(3) an NGO functioning as a trusted flagger under Article 22 of the DSA; and (4) an NGO functioning as trusted flagger outside of the DSA, in a private agreement with the internet intermediary. Since constitutional tensions arise in first and second scenarios, this chapter explores those further. The third and fourth scenarios create concerns on other levels, pertaining to the position of social media platforms as "enforcers" in the digital realm, which are well-discussed in literature (see, for example, Gillespie 2018; Kaesling 2018; Klonick 2018)—and on the level of legitimacy of the involvement of external parties in content moderation, for example, in the case of fact-checkers (Gillespie 2018) or external independent adjudicatory bodies (Klonick 2020).

In the first scenario, state actors are appointed as trusted flaggers by the DSC if they hold specific expertise and act diligently. Their flagging capabilities are limited to the constraints of "illegal content" under Article 3(h), and the form of Article 16. In theory, they can only flag content that is illegal under national or EU law; in practice, it is likely that trusted flaggers will continue to flag using community guidelines. However, a flag as laid down in Article 22 of the DSA fulfills the legality requirement of Article 10(2) of the ECHR: laws must be accessible and precise. Considering that national law must also be in accordance with EU law, this is unlikely to cause unlawfulness. There are two caveats to the requirement of lawfulness, however. Firstly, content can be flagged based on national law, making content illegal in one member state but not another. This decreases the legal certainty of internet users: it is excessive to require internet users to be acquainted with national law across the entire European Union. In this regard, geo-blocking has been an effective remedy (Lemley 2021): removing content only in regions where it is illegal overcomes issues with the lawfulness of that removal under Article 7 of the Gw and Article 10 of the ECHR. Secondly, the foreseeability of limitations to the right to freedom of expression in social media spaces is limited due to the opacity around content moderation remedies (Goldman 2021).

While terms of service agreements outline possible sanctions for violations of community guidelines, it often remains unclear which sanction is applied in a specific scenario. To address this lack of transparency, one solution is to enhance the clarity of the flags submitted by trusted flaggers within the notice and action mechanism. This could involve including an option for trusted flaggers to specify the remedy they are seeking. Furthermore, this information can be made available to the affected party, allowing them to see the internet intermediary's decision regarding the remedy based on the trusted flagger's referral.

If a referral restricts the right to freedom of expression, it must serve a legitimate aim under Article 10(2) of the ECHR. Legitimate aims can be found listed in that article and are interpreted broadly and against the cultural background of the state: what is deemed the protection of health and morals is not necessarily deemed so in other states. In notice and takedown mechanisms *ex* Article 16 of the DSA, it is common that the flagger can indicate law on which the flag is based. Transparency on the legitimacy of the restriction can be easily achieved by adding a choice menu to the flagging form, listing the legitimate aims of Article 10(2) of the ECHR. This creates transparency and accountability on the legitimacy of takedown requests; without such an indication, it is unclear whether takedown requests by trusted flaggers might interfere with the right to freedom of expression. Further, referrals potentially restricting freedom of expression must be necessary in a democratic society. This is a requirement of proportionality: the right must outweigh a pressing social need and be a suitable means to achieve this end. A proportionality assessment, explaining why the internet user's right to freedom of expression is outweighed by the societal need for removing his content, is currently lacking in content moderation and is not included in the statement of reasons *ex* Article 17 of the DSA. Including this in the statement of reasons, along with an explanation of why the chosen sanction is the suitable and necessary means to achieve the societal need it aims to address, decreases the risk for unlawful interferences with the right to freedom of expression.

A usual counterargument to the solutions proposed above is that individual rights-based approaches do not scale well, which is necessary in content moderation (Balkin 2018; Douek 2022; Sander 2020). However, since trusted flags concern individual cases, and the volume of trusted flags indicated by Dutch state organs is not such that individual case handling is impossible, it would be feasible to include such proportionality assessments in cases where a state body has functioned as a trusted flagger. This ensures that flags do not inadvertently violate the right to freedom of expression.

In the second scenario, state bodies function outside of the scope of the DSA in a private arrangement with the internet intermediary. This is the current practice. This enables state actors to flag content not only based on national or EU law but also based on the community guidelines of the social media space. This scenario gives rise to the same concerns as above but runs a further risk when it comes to the lawfulness of the flag. Eghbariah and Metwally describe the rule of law risk of referring based on community guidelines resulting in "state-interpreted service agreements" (2021). Presuming that a flag by a state body is a strong nudge

toward removal, and removal restricts the freedom of expression of internet users, it is problematic that such nudges can be made based on community guidelines. This is at odds with the requirement of lawfulness of Article 7 of the Gw and Article 10(2) of the ECHR. Further, this scenario has the opacity and legitimacy issues that the DSA has tried to overcome. Transparency and accountability are principles of good governance that can diminish in the existing trusted flagger framework for state bodies, in which they flag based on community guidelines. One way to overcome this is to only allow state bodies to function as trusted flaggers within the framework of the DSA: this ensures the lawfulness of their flagging and makes the extent of their flagging transparent. The downside is that this proverbially handcuffs state bodies in their quest to reduce societal risks caused by harmful content since they can no longer flag based on community guidelines. This could negatively affect the detection of "awful but lawful" content by internet intermediaries.

Conclusion

This contribution examined the trusted flagger framework in the Netherlands and the fragilities of the right of freedom of expression therein. Trusted flaggers are an exponent of networked governance that helps internet intermediaries engage with third parties' expertise in combating harmful content. Those third parties also involve state actors. Since a flag by a state actor functioning as a trusted flagger can be seen as a nudge toward removal of content, this can raise concerns for the protection of the freedom of expression. The Digital Services Act has attempted to legitimize the trusted flagger framework and remove the shroud of opacity that currently surrounds private arrangements between trusted flaggers and internet intermediaries. While it succeeds in some regards, it raises some concerns for the right to freedom of expression under Article 7 of the Gw and Article 10 of the ECHR when state entities operate as trusted flaggers, due to the indirect pressure for removal that might be exerted on the internet intermediary.

These concerns can be addressed with simple adjustments to the notice-and-action mechanisms used by internet intermediaries for trusted flaggers that better ensure the adherence to requirements for limitation of the freedom of expression laid down in Article 7(3) of the Gw and Article 10(2) of the ECHR. The lawfulness of flags can be ascertained by state actors solely flagging on the basis of national or EU law, by indicating the type of sanction they are looking for, and, if possible, by applying geo-blocking to

avoid unnecessarily blocking content in areas where it is not illegal. The legitimacy of those flags can be underlined by an indication of what aim it is serving under Article 10(2) of the ECHR. Since this is a finite list, adding one of the aims to a flag is not an excessive burden but does create transparency and accountability on the legitimacy of flags by state bodies. Finally, a flag by a state body should include an account of why the right of the internet user is outweighed by societies' needs, as well as an indication why the sought remedy is the appropriate way to fulfill those needs. Although this is not a scalable solution, it is possible to achieve this in the case-by-case context of trusted flagging. These are simple solutions to ensure that a valuable addition to the content moderation process—state bodies functioning as trusted flaggers—gains legitimacy and is ensured to respect the right to freedom of expression of internet users.

References

Appelman, Naomi, and Paddy Leerssen. 2022. "On Trusted Flaggers." *Yale Journal of Law & Technology* 24: 452–75. https://yjolt.org/trusted-flaggers.

Arcila, Beatriz Botero, and Rachel Griffin. 2023. "Social Media Platforms and Challenges for Democracy, Rule of Law and Fundamental Rights." PE 743.400. Policy Department for Citizen's Rights and Constitutional Affairs. European Parliament. https://sciencespo.hal.science/hal-04320778v1.

Are, Carolina, and Pam Briggs. 2023. "The Emotional and Financial Impact of De-platforming on Creators at the Margins." *Social Media + Society* 9(1): 1–12. https://doi.org/10.1177/20563051231155103.

Balkin, Jack M. 2018. "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation." *University of California, Davis Law Review* 51: 1149–10. https://doi.org/10.2139/ssrn.3038939.

Bambauer, Derek. 2015. "Against Jawboning." *Minnesota Law Review* 100(51): 51–128. https://www.minnesotalawreview.org/wp-content/uploads/2015/11/Bambauer_ONLINE.pdf.

Caplan, Robyn. 2023. "Networked Platform Governance: The Construction of the Democratic Platform." *International Journal of Communication* 17: 3451–72. https://ijoc.org/index.php/ijoc/article/view/20035.

Chang, Brian. 2017. "From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU." *Columbia Human Rights Law Review* 49(2): 114–212. https://hrlr.law.columbia.edu/files/2018/07/BrianChang-FromInternetRef.pdf.

Crawford, Kate, and Tarleton Gillespie. 2016. "What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18(3): 410–28. https://doi.org/10.1177/1461444814543163.

De Meij, Jan Marinus. 2000. *Uitingsvrijheid. De Vrije Informatiestroom in Grondwettelijk Perspectief.* Cramwinckel.

Douek, Evelyn. 2022. "Content Moderation as Systems Thinking." *Harvard Law Review* 136. https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/.

Eghbariah, Rabea, and Amre Metwally. 2021. "Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service." *Yale Journal of Law & Technology* 23: 542–617. https://yjolt.org/informal-governance-internet-referral-units-and-rise-state-interpretation-terms-service.

EU Code of Conduct. 2016. "EU Code of Conduct on Countering Illegal Hate Speech Online." https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_en.

European Commission. 2022. *The Strengthened Code of Practice on Disinformation.* https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven, CT: Yale University Press.

Goanta, Catalina. 2023. "The New Social Media: Contracts, Consumers, and Chaos." *Iowa Law Review Online* 108: 118–30. https://ilr.law.uiowa.edu/volume-108-response-pieces/2023/05/new-social-media-contracts-consumers-and-chaos.

Goanta, Catalina, and Jerry Spanakis. 2022. "Discussing the Legitimacy of Digital Market Surveillance." *Stanford Journal of Computational Antitrust* 2(April): 44–55. https://doi.org/10.51868/12.

Goldman, Eric. 2021. "Content Moderation Remedies." *Michigan Technology Law Review* 28(1): 1–60. https://doi.org/10.36645/mtlr.28.1.content.

Griffin, Rachel. 2023. "From Brand Safety to Suitability: Advertisers in Platform Governance." *Internet Policy Review* 12(3). https://doi.org/10.14763/2023.3.1716.

Hins, Aernout W. 1995. "Gedachten en Gevoelens over de Elektronische Snelweg." In *Communicatie- en Informatievrijheid in het Digitale Tijdperk*, edited by Jan W. Kalkman, Aernout W. Hins, and Erik C. M. Jurgens, 27–57. W. E. J. Tjeenk Willink.

Kaesling, Katharina. 2018. "Privatising Law Enforcement in Social Networks: A Comparative Model Analysis." *Erasmus Law Review* 11(3): 151–64. https://doi.org/10.5553/ELR.000115.

Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet.* New York: Columbia Global Reports.

Kilpatrick, Jane, and Chris Jones. 2022. "Empowering the Police, Removing Protections: The New Europol Regulation." Statewatch. https://www.statewatch.org/media/3615/empowering-the-police-removing-protections-new-europol-regulation.pdf.

Klonick, Kate. 2018. "The New Governors: The People, Rules, and Processes Governing Online Speech." *Harvard Law Review* 131(6): 1598–1670. https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf.

Klonick, Kate. 2020. "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression." *Yale Law Journal* 129(8): 2418–99. https://www.yalelawjournal.org/pdf/KlonickFeature_dsekuux4.pdf.

Kreimer, Seth F. 2006. "Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weaker Link." *University of Pennsylvania Law Review* 155(1): 11–102. https://scholarship.law.upenn.edu/penn_law_review/vol155/iss1/4/.

Lemley, Mark. 2021. "The Splinternet." *Duke Law Journal* 70(6): 1397–1428. https://scholarship.law.duke.edu/dlj/vol70/iss6/3/.

Mak, Vanessa. 2022. "Editorial: A Primavera for European Consumer Law: Re-birth of the Consumer Image in the Light of Digitalisation and Sustainability." *Journal of European Consumer and Market Law* 11(3): 77–80. https://kluwerlawonline.com/journalarticle/Journal+of+European+Consumer+and+Market+Law/11.3/EuCML2022014.

Sander, Barrie. 2020. "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights–Based Approach to Content Moderation." *Fordham International Law Journal* 43(4): 939–1006. https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3/.

Teubner, Gunther. 2017. "Horizontal Effects of Constitutional Rights in the Internet: A Legal Case on the Digital Constitution." *The Italian Law Journal* 3(1): 193–208. https://theitalianlawjournal.it/index.php?id=teubner-1.

Tweede Kamer der Staten-Generaal. 2023. "Kamerstukken 2022–2023, Nr. 1599." February 20. http://zoek.officielebekendmakingen.nl/ah-tk-20222023-1599.html.

Urban, Jennifer M., Brianna L. Schofield, and Joe Karaganis. 2017. "Takedown in Two Worlds: An Empirical Analysis." *Journal of the Copyright Society of the USA* 64(4): 483–520. https://doi.org/10.31235/osf.io/mduyn.

About the Author

Jacob van de Kerkhof is a PhD candidate with the Montaigne Centre at Utrecht University.