van de Kerkhof, Jacob

Article

Article 22 Digital Services Act: Building trust with trusted flaggers

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: van de Kerkhof, Jacob (2025) : Article 22 Digital Services Act: Building trust with trusted flaggers, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 1, pp. 1-26, https://doi.org/10.14763/2025.1.1828

This Version is available at:
https://hdl.handle.net/10419/315582

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 14 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Article 22 Digital Services Act: Building trust with trusted flaggers

Jacob van de Kerkhof *Utrecht University*

DOI: https://doi.org/10.14763/2025.1.1828

Published: 31 March 2025
Received: 5 April 2024 **Accepted:** 21 October 2024

Funding: The authors did not receive any funding for this research.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: van de Kerkhof, J. (2025). Article 22 Digital Services Act: Building trust with trusted flaggers. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1828

Keywords: Content moderation, Digital Service Act, Freedom of expression, Trusted flaggers

Abstract: Trusted flaggers have long played a role in content moderation: a bilateral, voluntary affair between online platforms and individuals or organisations that are afforded prioritised access to the content moderation process. Due to their expertise, they are trusted to 'flag' illegal or harmful content. Article 22 Digital Services Act formalises this framework, allowing governmental and non-governmental organisations to apply for certification as trusted flaggers and requiring online platforms to treat their submitted notices on illegal content with priority and without undue delay. The certification of the first trusted flaggers under Article 22 has sparked public debate about their influence and power, especially in relation to the freedom of expression of internet users. Concerns about trusted flagger frameworks are new in part, but also reflect existing weaknesses in the framework relating to over removal and freedom of expression. This contribution explains the concerns about trusted flagger involvement in content moderation in light of freedom of expression, assesses the promises and pitfalls of Article 22 in its current implementation, and offers recommendations to ensure more effective operationalisation of trusted flaggers under Article 22 and better safeguard the right to freedom of expression of internet users.

This paper is part of **Content moderation on digital platforms: beyond states and firms**, a special issue of *Internet Policy Review* guest-edited by Romain Badouard and Anne Bellon.

Introduction

The selection of the first trusted flaggers under Article 22 Digital Services Act (Regulation 2022/2065, 'DSA') has sparked public debate, fuelled by concerns over their power and potential impact on freedom of expression on the internet (*e.g.* Schneider, 2024). Trusted flaggers are individuals or organisations that are afforded prioritised access to the content moderation process of online platforms. Commonly, they have access to a privileged channel to 'flag' content – submit a notice indicating illegality or incompatibility with community guidelines (Crawford & Gillespie, 2016) - which is subsequently addressed with priority by the online platform compared to the treatment of 'ordinary' flags. Access to this privileged flagging mechanism is based on expertise in content areas. Organisations functioning as trusted flaggers range from national enforcement authorities, such as the Dutch Ministry of Internal Affairs (*Kamerstukken* 2022-2023, nr. 1599), to non-governmental organisations (NGOs), such as the INHOPE network reporting child sexual abuse material (CSAM). Trusted flaggers are commonly seen as positive additions to the content moderation process, but have remained relatively understudied actors (Schwemer, 2019). While trusted flaggers act primarily in their own interest, seeking to combat illegal or harmful content (Appelman & Leerssen, 2022), they also support the interests of online platforms, who want to rid their platforms of harmful content to make their space safer and more attractive for users and – crucially – advertisers (Griffin, 2023a). However, the role of trusted flaggers in content moderation has also been subject to critique. Namely, that trusted flaggers can be 'censoring' through opaque means, potentially pushing a political narrative and controlling the public discourse by removing unwanted content (e.g. Veil, 2024). This is amplified when a state entity as a trusted flagger submits notices that lead to removal because it implies sidestepping constitutional norms to unduly interfere with freedom of expression (van de Kerkhof, forthcoming).

Originally, trusted flagger arrangements were a voluntary effort by online platforms to channel the expertise of civil society into their content moderation process. Examples of such programmes include Amazon's Project Zero, Meta's Trusted Flagger programme, and YouTube's Priority Flagger programme. The EU-regulator has adopted this idea in co-regulatory documents such as the Strengthened Code of Practice on Disinformation 2022 (comm. 21) and the Code of Conduct

on Countering Illegal Hate Speech Online 2016, encouraging providers of online platforms to offer voluntary trusted flagger frameworks. Similar examples are found on a national level. For example, the now-defunct 2016 German *Netzwerkdurchsetzungsgesetz* created trusted flagger-like entities called *Beschwerdestellen:* expert organisations that forward complaints to providers of online social media platforms, such as *Jugendschutz.net*.

The DSA codifies this formerly voluntary framework in Article 22. The DSA-enforcing authorities at member state-level - Digital Services Coordinators (DSCs) - select trusted flaggers that must be recognised by all providers of online platforms (Article 22(1)). Once the trusted flagger is awarded their status, providers of online platforms must facilitate notice submissions through the mechanisms created in line with Article 16 DSA, and ensure these are given priority and processed without undue delay. Article 16 DSA requires providers of online platforms to create notice and action mechanisms that allow users to notify them about illegal content, providing an explanation of why the individual thinks the information is illegal, where it can be found, the identity of the user, and a statement of good faith. The contribution of this paper is to (1) situate Article 22 DSA against the background of current critiques of trusted flagger frameworks in light of the protection of freedom of expression in the European Union, (2) identify promises and pitfalls in the new framework offered by Article 22, and (3) propose steps that can be taken to provide better safeguards regarding freedom of expression, more harmonised implementation of Article 22 and better operationalise the trusted flagger framework. The article is structured as follows: Section I highlights existing critiques of the trusted flagger framework; Section II situates Article 22 against that backdrop and assesses whether it has delivered on promises of legitimacy or raises additional pitfalls; Section III finally proposes concrete measures that can be taken to ensure legitimate and harmonised implementation of Article 22 DSA. It is currently early days regarding the implementation of Article 22: as of January 2025, there are only 16 selected trusted flaggers. This article therefore makes a timely and novel contribution by (a) clarifying the current public debate on trusted flaggers in the DSA context, and (b) proposing steps that can be taken by the European Commission to harmonise the application of Article 22 under Article 22(8) - creating the possibility for the Commission to issue further guidance-, improve freedom of expression safeguards, and create a more effective trusted flagger framework.

The article draws on legal and socio-legal sources related to trusted flaggers. The article builds on case law from jurisdictions beyond the EU, by lack of EU case law, to illustrate what tensions may arise within the EU. In terms of the fundamental

rights framework drawn upon, this article refers to EU fundamental rights as expressed in the Charter of Fundamental Rights ('CFR'), which can be interpreted in light of case law by the European Court of Human Rights ('ECtHR') following article 52(3) CFR.

Section I: The existing trusted flagger framework and its weaknesses

The existing trusted flagger framework in content moderation is a private initiative, relying on the willingness of online platforms to engage with individuals or organisations with particular expertise to aid in content moderation. In doing so, online platforms provide trusted flaggers with certain privileges in the content moderation process, ranging from dedicated flagging portals and employees to whom complaints can be submitted to holding an advisory role in setting community guidelines. This allows trusted flaggers to represent particular interests that might otherwise be underrepresented, such as those of minorities or vulnerable groups (Appelman & Leerssen, 2022, p. 467). One example of such groups is the INHOPE network, focussing on combating CSAM content. Although the trusted flagger framework is generally perceived as a positive step toward more effective and legitimate content moderation due to the expertise of the flagger and the potential for civil-society involvement, scholars have also criticised the framework. This section outlines those critiques, against which Article 22 DSA is subsequently evaluated in Section II. The critique centres on three points: (i) risk of over-removal; (ii) fundamental rights and rule of law concerns; and (iii) transparency.

(i) Risk of over-removal

One of the primary concerns in the landscape of content moderation is the risk of over-removal. It has been theorised that internet liability laws would lead platforms to excessively remove content because they would likely 'err on the side of caution' in their moderation practices in order to avoid liability (Chang, 2018, pp. 122–123). The same argument is levelled at trusted flaggers. A trusted flagger is afforded flagging privileges on the basis of their expertise (Kaesling, 2022, p. 341) and the quality of their work (Husovec, 2024, p. 207). Therefore, their flags might carry more weight, creating a risk for what Schwemer calls 'rubberstamping': a situation in which the online platform restricts content without in-depth review (Schwemer, 2019, p. 8). This is in line with the European Commission's attitude to trusted flaggers, stating that every flag should be treated with "an appropriate degree of confidence as regards their accuracy" (Recommendation 2018/334, rec. 29). Limited platform review becomes problematic when a trusted flagger errs about il-

legality or incompatibility with terms of service. In such cases, trusted flagger involvement could unduly restrict internet users' speech, impeding their freedom of expression, which is further discussed in the following section. Appelman and Leerssen suggest that this risk for over-removal is driven by the fact that trusted flagger organisations often have specific interests, and they can therefore be aggressive in submitting notices pertaining to content conflicting with those interests (Appelman & Leerssen, 2022, p. 468). In doing so, their voices can outweigh those that propose a more balanced approach with regard to freedom of expression (Dvoskin, 2022, pp. 469-475). Platforms are expected to comply with these flags under the DSA risking over-removal of potentially legal speech.

Empirically, the risk for over-removal is difficult to assess, since online platforms are rarely transparent about the specific content they remove, and information about the proportion of notices submitted that led to removal is even more opaque. Some circumstantial evidence can be found in literature: Eifert and colleagues' evaluation of the *NetzDG* shows that complaints by complaint bodies (*Beschwerdestellen)* were removed at a marginally higher rate (18% vs 14%) than complaints by individuals (Eifert et al., 2020, p. 40). This indicates that, although such notices are more likely to be removed, this does not happen to the extent that it should be considered over-removal or censorship. A slight increase in removal rates might be expected because of the expertise of trusted flaggers in identifying illegal or incompatible content. On the other hand, Urban and colleagues' 2017 study shows that copyright holders submitting notices through privileged avenues has led to a significantly higher rate of removal, and that removal often happened without significant review on the part of the online platform (Urban et al., 2017, pp. 54–55). This could indicate over-removal, which would be problematic: the interest of the IP-holder is not the protection of freedom of expression, but a commercially-incentivised protection of copyrighted creations, which might inspire erring on the side of caution in submitting notices, in line with Dvoskin's observation that interests groups advocating more content moderation have outweighed those advocating less. That said, with some exceptions, the illegality of IP-infringing content can be less contentious from a freedom of expression perspective; in cases of piracy or counterfeiting, content flagged can be clearly illegal, whereas in cases of political speech, this can be more complex as hate speech and misinformation are more gray area policy fields.

Some data is also available on the Dutch Ministry of the Interior acting as a trusted flagger with Facebook and Twitter (van de Kerkhof, forthcoming). In parliamentary debate, excerpts of takedown requests were shared: the Ministry was flagging

disinformation around elections on behalf of municipalities who struggled to gain access to social media platform operators (*Kamerstukken* 2022-2023, nr. 914). Misinformation is a legally contentious issue, and the border between legal and illegal misinformation is often unclear. Facebook responded accordingly to this ambiguity, not complying with their flags on disinformation. However, upon government requests about reinstating blocked accounts, Facebook reinstated accounts of government officials and posts thus complying with the Ministry's wishes (*Kamerstukken* 2022-2023, nr. 914). While the sample size of these requests is very small (5), it illustrates that in legally contentious areas, platform providers can be more careful with 'rubberstamping', pushing back to government wishes. This suggests that platform review can in potential be an effective safeguard in the trusted flagger framework, and 'rubberstamping' could be overstated. It is difficult to assess whether the trusted flagger framework really leads to over-removal – likely this is dependent on the actor requesting, the nature of the content on which the request is submitted and the incentive for the platform to comply. Even if content is not over-removed, trusted flagger practices can still lead to fundamental right concerns, as explained in the following paragraph.

(ii) Fundamental rights and rule of law concerns

The right to freedom of expression of internet users can be at risk when legal content is removed without justification. In the European Union, freedom of expression is protected under article 11 CFR, which, following article 52(3) CFR, can be interpreted using case law on article 10 of the European Convention on Human Rights (ECHR) of the European Court of Human Rights, since those articles cover the same right. Freedom of expression covers the right to express yourself on the internet (e.g. *Melike v Turkey,* 35786/19 (2021), ECtHR), and the right to information (*Cetin et al v Turkey,* 40153/98 & 40160/98 (2003), ECtHR, para 64). Content being taken down can interfere with that right, either by limiting the speech of the internet user or the right to information of the internet user who can no longer access that content. Interferences to these rights are allowed, if they meet the requirements of article 52(1) CFR, meaning they pursue a legitimate aim, are prescribed by law, and are necessary in a democratic society (*The Sunday Times v UK,* 6538/74 (1979), ECtHR, para 49; *Animal Defenders v UK*, 48876/08 (2013), ECtHR, para 100). Usually, such interferences are subject to judicial review, assessing if the listed requirements are met; when submitting informal requests – 'flags' – such balancing acts are not made by a judicial authority, but by the entity submitting the notice, and ultimately by the platform moderating content. Although the final balancing act is indeed made by the platform, it is possible that pressure from such requests by privileged entities leads to removal (Chang, 2018, pp. 122–123). This creates a

scenario in which although the platform makes a voluntary decision to remove content, the pressure exuding from a removal request resulting in removal can interfere with the internet user's freedom of expression.

The tension with fundamental rights is present in the case of private entities acting as trusted flaggers as well as government entities acting as trusted flaggers. Private entities are not directly bound by fundamental rights norms (De Gregorio, 2022, pp. 196–200; Haupt, 2024, p. 10), but have a responsibility, implicitly in EU national case law (e.g. *Facebook,* III ZR 179/20, 2021, BGH) and explicitly in the DSA, e.g. Article 14(4), to respect fundamental rights norms in the provision of their services. For state entities, acting as a trusted flagger can create a more pertinent fundamental right tension (van de Kerkhof, forthcoming). State entities carry particular responsibility with regards to the freedom of expression of internet users; more so than platforms. State entities issuing informal removal requests can interfere with freedom of expression (van de Kerkhof, 2024), since their notices likely carry more weight due to the potential regulatory sanctions that can follow from non-compliance.

In various non-EU jurisdictions, courts have deliberated the constitutionality of state entities functioning as trusted flaggers with regards to freedom of expression, illustrating the constitutional strain occurring when governmental authorities engage – often in an opaque, non-accountable manner – in content moderation. For example, the Israeli Supreme Court addressed questions about state entities' privileged access to content moderation in Adalah Legal Center v State Attorney's Office. In that case, Adalah had complained about Palestinian freedom of expression being violated due to social media platforms being pressured by flags by the Israeli police cyber department. The Israeli High Court found that such requests do constitute a state act that requires a legal basis, since state actors are repeat players that have coercive regulation available to them (Adalah Legal Center for Arab Minority Rights in Israel v State Attorney's Office, Cyber Department, HCJ 7846/19 (Isr.)). However, as long as it is not shown that a flag by a state agent causes the platform to forgo independent assessment or that the flag impedes a platform's discretion, there is not a direct violation of freedom of expression, the High Court concluded.

In the US, the Supreme Court decided on a similar matter in Murthy v Missouri. The case considers dealings between White House officials and social media platforms requesting takedown of Covid-19 misinformation. The 5th Circuit of Appeals had decided that a platform's content moderation practices can be so induced by government conduct that there is coercion or significant encouragement that such

conduct can be seen as a First Amendment violation (Missouri v Biden, 595 U.S. ___ (2022), 5th. Cir., p. 50-51). The Biden Administration had interacted with platforms in a manner that was 'coercive', or at least 'significantly encouraging', affecting the independence of the platform. The Supreme Court did not address the alleged coercion substantively, rather finding that the respondents had no standing to bring suit (Murthy v Missouri, 601 U.S. ___, (2024),S. Ct.). However, in his dissent, Justice Alito indicated that he thought that the requests were unconstitutional (Murthy v Missouri, 601 U.S. ___, (2024),S. Ct., Alito dissenting, p. 16). Social media platforms are also acutely aware of the tension arising from government authorities acting as trusted flaggers.

In UK Drill Music, the Meta Oversight Board emphasised that law enforcement can provide important context and expertise in the content moderation process, but it is important that Meta platforms make independent assessments based on that information and accurately represents its engagements with law enforcement on these matters (UK Drill Music, Oversight Board, 2022). The cases above illustrate that government authorities engaging with content moderation can cause a tension with fundamental rights; however, in the EU, courts are yet to deliberate the freedom of expression concerns associated with the use of trusted flaggers.

Aside from concerns for freedom of expression, in cases where state entities function as trusted flaggers, there can also be concerns for the rule of law, in light of their responsibility to act in accordance with the law. Content moderation is based either on the law or on the basis of platforms' community guidelines. Acting as a trusted flagger provides state entities a privileged channel to submit notices on the basis of community guidelines, creating a rule of law tension: *de facto* state entities could be enforcing private guidelines, not the law. Under the rule of law, state entities require a legal basis to engage in content moderation, even if it simply means encouraging platforms to enforce their own community guidelines (Schwemer, 2019, p. 10). That legal basis is missing when content is moderated on the basis of its terms of service. It is well-documented that terms of service are the preferred basis of content moderation (e.g. Elkin-Koren et al., 2022; Quintais et al., 2023). Empirically, researchers have confirmed this by studying the DSA Transparency Database holding statements of reasons on content moderation decisions by online platforms (Kaushal et al., 2024; Trujillo et al., 2023). UN Special Rapporteur Kaye provides two reasons why this practice could be problematic when state entities are involved: (i) terms of service are often broadly phrased (see also Citron, 2018, p. 1058), meaning that there is little legal certainty of what content is regarded as illegal, and (ii) terms of service agreements often go beyond what is pro-

hibited by law (Kaye, 2016, pp. 14–15). State entities using terms of service to flag content can *de facto* sidestep legal requirements to request content removal. Eghbariah and Metwally flag this issue, citing instances of police internet referral units – police departments that focus on investigating malicious content on the internet - wrongly targeting extremist content online relying on broad definitions in terms of services resulting in unwarranted censorship (Eghbariah & Metwally, 2021, pp. 587–589). This creates a paradox: states and platforms rely on definitional ambiguity in terms of service and lack of scrutiny present in trusted flagger frameworks because they are a low-cost, highly expedient way of targeting not only unlawful, but also legal but harmful content on the internet, which may ultimately benefit society as a whole, but comes at the expense of strict adherence to fundamental rights safeguards (Schwemer, 2019, p. 12).

(iii) Transparency

The paragraphs above have alluded to the current trusted flagger framework being opaque. (Lack of) transparency is an oft-discussed topic in content moderation. The general consensus is that content moderation is an opaque process, increasing the accountability deficit in regulating online platforms (Suzor, 2019; Leerssen, 2020, 2024; Gill, 2021; Rieder & Hofmann, 2020). Increasing transparency about content moderation practices is suggested to potentially increase accountability in the content moderation process (Bloch-Wehba, 2019; Griffin, 2023b; Kaye, 2019), but the operationalisation of transparency has inspired critique on allowing online platforms to control the discourse over their practices (Flyverbom, 2016), yielding inconsistent results (Kosta & Brewczyńska, 2019; Urman & Makhortykh, 2023), and suffering from a general lack of applicability to the process of algorithmic content moderation due to the opaque nature of automated moderation (Ananny & Crawford, 2018; Edwards & Veale, 2017).

Private arrangements of platforms with entities operating as trusted flaggers are similarly opaque. Platforms do not need to be forward about whom they afford flagging privileges, leaving users unable to control what actors are involved in content moderation. Platforms, meanwhile, control the discourse about their trusted flagger framework by being forward about their interactions with NGOs, but not about their interactions with governmental organisations or IP-right organisations, which are generally less popular (Appelman & Leerssen, 2022, pp. 470–471). In practice, this is confirmed in the Oversight Board decision *UK Drill Music*. The Oversight Board found that transparency reports underrepresent interactions with law enforcement, raising concern since trusted flaggers can hold a flagging bias to which feedback is not available. This relates to submissions made by trusted flag-

gers, but also to trusted flagger practices as a whole. The funding, organisational structure, and (in)dependency of trusted flaggers cannot be assessed if it is unknown who these organisations are, leaving little room for accountability. This lack of transparency creates an accountability deficit, which can harm the legitimacy of the trusted flagger framework. Schwemer, building on Suchman's approach to legitimacy as 'a generalised perception of desirability' (Suchman, 1995), theorises that legitimacy is also "about being representative" of a group, the general public, or a specific interest (Schwemer, 2019, p. 8). In the current opaque scenario, there is no way of holding organisations accountable for the desirability of their actions, practices, and the interests they represent, meaning that this detracts from the legitimacy of the trusted flagger framework.

Section II: Article 22 DSA

This section positions the framework of Article 22 against the backdrop of current critiques of trusted flagger frameworks, identifying the promises and pitfalls of the DSA trusted flagger system. The DSA is a newly entered-into-force EU Regulation that aims – in part – to ensure better protection of fundamental rights online. Article 22 DSA establishes that providers of online platforms must take measures to ensure that trusted flaggers, as appointed by the DSCs, can submit notices through the mechanisms referred to in Article 16 DSA, and those notices are processed with priority and without undue delay. Article 22 departs from private trusted flagger arrangements in which online platforms bilaterally engage with trusted flaggers, and forms a logical next step after a series of nudges in the European Union that stimulate online platforms to include trusted flaggers in the content moderation process (Husovec, 2024, p. 207; Schwemer, 2019, pp. 10–11). The framework creates a formal avenue for trusted flaggers to gain privileged access to content moderation with *all* providers of online platforms active in the EU. This can benefit smaller organisations that have struggled to develop a relationship and gain access to a platform, but also help larger organisations gain access to platforms that currently do not have a trusted flagger framework, for example X (Dang, 2022). The compulsory nature of the Article 22 trusted flagger framework is particularly relevant in a landscape where companies such as Alphabet and Meta are increasingly withdrawing from content moderation efforts in search of political favour (van de Kerkhof, 2025). Regulatory pressure ensures the continued participation of selected trusted flaggers in the content moderation process.

The DSC appoints trusted flaggers based on their (i) particular expertise (ii) independence from any platform provider and (iii) diligence, accuracy and objectivity

(DSA, art. 22(2)). Trusted flaggers can be governmental and non-governmental organisations, but not individuals (DSA, rec. 61). DSCs must avoid appointing too many trusted flaggers, which could decrease the value of the framework, hence precluding individual rights holders from applying (Raue, 2023, p. 398). Trusted flaggers must publish yearly reports with the DSC of the notices they have submitted with platform providers. In the event that a trusted flagger submits a significant number of imprecise, inaccurate or unsubstantiated notices, the DSC may, after being informed by the provider of an online platform, open investigations on that trusted flagger and simultaneously suspend the organisation. Investigation can lead to revocation of trusted flagger status. It is important to note that Article 22 is without prejudice to existing trusted flagger arrangements (DSA, rec. 62): existing arrangements can continue to exist. This can create a tension that is addressed in subsection II.ii.

(i) Risk of over-removal

Article 22 DSA forms both a solution and exacerbator to the risk of over-removal as a consequence of the trusted flagger system, as discussed in the previous section. Because Article 22 grants certified trusted flaggers access to *all* online platforms, trusted flaggers with sufficient resources have a wider range of operations than they previously had – potentially enlarging the risk for over-removal. However, the DSA mitigates the risk of over-removal in the trusted flagger framework at four instances: the trusted flagger, online platforms, DSC, and user.

Firstly, the trusted flagger has to report on its activities yearly under Article 22(3), including the number of notices accompanied by the type of illegal content and the action taken by the provider. There is accountability in transparency: transparency reports will show if a trusted flagger overextends its competence by submitting notices that do not address illegal content in the sense of article 3(h), and thus do not require action by online platforms. Over-removal is mitigated further by the requirement for accuracy and expertise in the certification of trusted flaggers: DSCs have to select organisations that have demonstrable expertise in their field, and therefore certified organisations are less likely to submit incorrect notices.

Secondly, online platforms also have a defence against trusted flaggers over-submitting notices: Article 23(2) DSA gives online platforms the right to suspend flagging privileges if entities frequently submit notices that are manifestly unfounded – this includes trusted flaggers (Raue 2023, p. 405). Similarly, if notices are structurally unfounded because they are overturned in internal appeal procedures un-

der Article 20(4) DSA online platforms can request the DSC to suspend a trusted flagger under Article 22(6) DSA. Thirdly, the DSC can revoke the status of a trusted flagger if it is over-submitting under Article 22(7) DSA. Fourthly, the user of the online platform is provided with a range of possibilities for appeal if they find that their content is wrongly moderated in Article 20 and Article 21 DSA. This, in a general sense, mitigates the risk for over-removal: in the previous bilateral trusted flagger structure, the user had no means to revert a content moderation decision – the DSA provides those means, thus somewhat countervailing the power of the trusted flagger.

Section I.i proposes that the risk of over-removal arises from trusted flaggers representing specific interests, leading to aggressive flagging practices that do not necessarily balance freedom of expression with the interest they represent. This risk can be present in the interpretation of the selection criteria for trusted flaggers under Article 22 DSA. The selection criteria of Article 22(2) DSA are multi-interpretable, and can be coloured based on national preference, since trusted flaggers are ultimately certified by national DSCs. In an EU landscape where rule of law backsliding exists (Scheppele, 2023) and attitudes toward speech vary regionally, it is possible that organisations that are selected could form a risk to freedom of expression (Orlando-Salling & Bartolo, 2023). Such organisations could potentially over-submit notices with online platforms, leading to over-removal. The selection criteria as they are laid down in article 22 do not provide an effective backstop to this: trusted flaggers are selected on particular expertise (Art. 22(2)(a)), implying they will have a particular interest in combatting a specific type of illegal content. Article 22(2)(c) then specifies that the trusted flagger must be able to carry out its activities diligently, accurately and objectively. '*Objectively*' as a requirement could form a safeguard to ensure that the trusted flagger weighs their interests with that of the freedom of expression of internet users. However, that criterion is, in the selection procedure, generally demonstrated through submitting a range of successful notices as evidence (Raue, 2023, p. 399). Unsuccessful notices would generally be flagged by the platform, who, as explained above, has little interest in disagreeing with the trusted flagger's notices. It is therefore difficult to gauge whether the requirement of 'objectivity' *ex* Article 22(2)(c) provides a sufficient counterbalance to the explicit interests that trusted flaggers have that could result in over-removal.

Next to the multi-interpretable norms for selection, the DSA does not specify the means trusted flaggers must use to detect content. While some trusted flaggers rely on hotlines or their own investigation representing a more artisanal, case-by-

case approach to content moderation (Caplan, 2018), others may look to use algorithmic means to detect illegal content: this is common practice for entities detecting CSAM or copyright infringing content. The risks for freedom of expression related to algorithmic content moderation have been well documented (*e.g.* Casarosa, 2021; Castets-Renardt, 2020; Gorwa et al., 2020). The use of algorithmic content moderation for detection by trusted flaggers raises the same concerns: is it desirable that the scalability of their activities is achieved at the expense of accuracy? Using automated means also raises a transparency concern: online platforms are required to indicate whether they use automated means to detect illegal content and whether they rely on automated decision-making in the statement of reasons they send to users under Article 17 DSA, but if the flag is submitted by a trusted flagger the user is not informed of any automated detection mechanisms used, *de facto* sidestepping the requirement of Article 17 DSA.

Aside from remedies pertaining to specific removal of content, the user has no means to challenge the privileged status of a trusted flagger. Trusted flaggers are selected by the DSC, a process which is subject to national administrative law. Whether it is possible for interested parties to appeal such a decision depends on the administrative law of the member state, fragmenting the degree to which users are able to have redress against the selection of trusted flaggers. This, coupled with the fact that trusted flagger statuses can only be revoked by the DSC that awarded the status, makes it challenging for users who seek to challenge a trusted flagger established in a different member state.

(ii) Fundamental rights and rule of law concerns

Section I.ii outlines the fundamental rights risks of the existing trusted flagger framework. Those risks predominantly pertain to state entities flagging as trusted flaggers, as their interference with freedom of expression without providing sufficient justification under article 52(1) CFR is problematic from a fundamental rights standpoint. The concern for fundamental rights is accompanied by rule of law concerns related to flagging on the basis of terms of service. Article 22 DSA reduces both risks, but there are some caveats.

First of all, the number of state entities seeking certification under Article 22 is tempered somewhat by Article 9 DSA, which provides a clear avenue for administrative and judicial authorities seeking to order the takedown of illegal content – they can do so directly with the platform with more legal safeguards in place – as well as the sentiment expressed in DSA recital 61 suggesting that the number of trusted flaggers should be kept to a minimum. The primary concern in Section I.ii

was that state entities acting as trusted flaggers can submit notices that carry enough weight for platforms to remove content without proper review. By doing so, the state entity is able to indirectly remove content without providing a legal basis, a legitimate aim, and weighing the necessity in a democratic society – which are the requirements for interfering with the freedom of expression. Under Article 22, this risk is mitigated: trusted flaggers can only flag content through the mechanisms of Article 16 DSA. Article 16 DSA requires that notices must be submitted on the basis of *illegal content,* which is defined in article 3(h) DSA as "*any information that […] is not in compliance with Union law or the law of any Member State which is in compliance with Union law […].*" This provides a degree of legal certainty to users, and decreases the concern of state entities sidestepping the requirements of 52(1) CFR and article 10(2) ECHR: certified trusted flaggers can only submit notices on content that is illegal – and not when it is non-compliant with terms of service.

Despite the additional layer of accountability, there are three loopholes to the operationalisation of flagging through Article 16 that form a risk to the rule of law, not only for state entities but for all trusted flaggers. Firstly, although Article 16 specifies that notices can only be submitted for illegal content, appointed trusted flaggers could still *de facto* flag on the basis of terms of service, just not within the framework of Article 16. If a notice is submitted outside of Article 16, there is no need for transparency reporting on such notices under Article 22(3). Conversely, it is not expected that online platforms provide the same priority and expedience to these notices. However, Recital 62 hints that online platforms should provide similar (priority) treatment to notices submitted by entities that are not awarded trusted flagger status to "take quick and reliable action against content that is incompatible with their terms and conditions". It is unclear how this sentiment relates to the requirement for priority under Article 22 – after all, priority can only be given once (Raue, 2023) - but it shows that trusted flagger mechanisms can *de facto* be used to flag on the basis of terms and conditions. Only *de jure*, it cannot under Article 22 DSA.

Secondly, the expected implementation by online platforms of Article 22 raises concerns: some platforms that have a framework will simply onboard newly certified trusted flaggers into the existing framework. Amazon, Facebook and Instagram have pledged that their companies will include new trusted flaggers within their existing ecosystems (*Conference's notes - Getting Ready for the DSA*, 2024), which allows for flagging under terms of services under the standards of Article 14 DSA. This is somewhat in the spirit of Recital 62, suggesting that trusted flaggers out-

side of Article 22 DSA can be treated similarly as those certified by a DSC. However, such portals usually allow for flagging both on the basis of terms of service as well as the law, maintaining the risk for the rule of law.

A third loophole is found in the backstop of Article 20(4) DSA, which serves as a requirement for DSCs to start investigations on trusted flaggers per Article 22(6) DSA. Article 20 describes the DSA's internal complaint handling system: when a user finds that their content is wrongly restricted, they can submit a complaint through the online platform's complaint handling system. If the platform finds that the content is not illegal and not incompatible with the platform's terms and conditions, it reverses its content moderation decision. However, it may be that the flagged content is not illegal but is incompatible with terms and conditions. For example, consider the case when a trusted flagger has submitted a notice per Article 16 on content it considers hate speech, which the platform acts upon and restricts the user's content. When the user submits a complaint through Article 20, and the platform reassesses its content moderation decision, it finds the content was not illegal under national hate speech law, but is still incompatible with the platform's terms and conditions. The content in that case is still removed, but the legal basis for its removal has changed – it is no longer removed on the basis of the law, but solely on the basis of terms of service, thus creating tension with the rule of law safeguard anchored in illegal content in Article 16.

Another point under the DSA trusted flagger framework that forms a challenge to the rule of law is the transnational nature of the internet. Content that is illegal in one member state can be legal in another, but trusted flaggers combatting illegal content receive their status for the entire EU. This can lead to the removal of content that is potentially legal in a given member state but illegal in another (Mauritz, 2024). For example: LGBT+ related content in advertisements is illegal in Hungary (Act LXXIX of 2021), but not in other EU member states (*European Commission v Hungary*, C-769/22 (2022), CJEU). A Hungarian trusted flagger could flag this content outside of Hungary, leading to its removal by lack of a vigilant online platform, even though the legal basis in that member state is missing. The exercise of determining whether the legal basis for the notice submitted is valid in another member state is left to the discretion of the trusted flagger and the discretion of the platform, who ultimately decides whether the content is restricted (Roundtable on the Digital Services Act, 2024). Both are not necessarily equipped to check whether the legal basis for removal contravenes the sentiment of Article 3(h) DSA that national law must be in compliance with EU law before content is actually illegal. Therefore, the legal certainty of users can be at risk when trusted flaggers

can *de facto* start enforcing their national law across other member states. Platforms would be encouraged to use geo-blocking as an adequate remedy, ensuring that content remains available where it is legal (Lemley, 2021).

(iii) Transparency

The previous section has explained that the current trusted flagger framework suffers from an opacity problem. As hinted above, Article 22 DSA provides improvements in that regard. A primary benefit lies in knowing what organisations have trusted flagger access to content moderation processes: the European Commission maintains a database with accredited trusted flaggers (Art. 22(5) DSA). This allows the identification of trusted flagger organisations, which adds to their legitimacy as part of Schwemer's representation perspective. Additionally, trusted flaggers need to publish reports with the DSC on notices they submit to online platforms in a yearly report. That report shows the number of notices, the nature of the content reported, and the action taken by the platform following the notice (Art. 22(3) DSA). This enables scrutiny over trusted flaggers' activities: do they flag content within their mandate based on their expertise? Are they over/under submitting? Scrutiny can ultimately lead to suspension or revoking of trusted flagger status under Article 22(6) and (7). Transparency reporting also enables scrutiny over the platform's response to trusted flagger activities: the remedy attached to the notice, as well as the response time, which indicate how well platforms comply with Article 22 DSA. The possible scrutiny of trusted flagger practices also adds to the legitimacy of the trusted flagger framework, since, in line with Suchman's notion of legitimacy, this allows the assessment of the desirability of their actions.

The DSA's intention is for trusted flaggers to positively influence content moderation by ensuring that they help platforms take action against illegal content more quickly and reliably (rec. 61 DSA). That intention is dependent on the goodwill of platforms to engage with those trusted flaggers; if online platforms do not respond to trusted flaggers, content will remain unchallenged. Transparency reporting by trusted flaggers on how platforms respond to the notices they have submitted allows DSCs to scrutinise platforms and potentially start an investigation or infringement procedures, thus forming a crucial redress instrument for trusted flaggers against uncooperative online platforms.

However, the DSA has not required providers of online platforms to share in their statement of reasons provided to users whether their restricted content was flagged by a trusted flagger (Art. 17(3)(b) DSA). Therefore, direct accountability of a trusted flagger by the general public is difficult: they will not know whether their

content is flagged, and, therefore, are unable to hold the trusted flagger to account. This makes it impossible to privately enforce the DSA-rights and even GDPR-rights (a point beyond the scope of this article, but trusted flaggers can qualify as data processors) against trusted flaggers. This also creates a caveat in the liability for wrongful moderation by online platforms. In principle, one can hold platforms liable for wrongfully moderating content, either through out-of-court dispute settlement bodies *ex* Article 21 DSA, or through a national court. Online platforms have plenty of incentive to abide by requests of trusted flaggers: they can rely on the expertise of specialised entities which benefit from the legitimacy of being selected by national authorities, a deficit of the trusted flagger framework before Article 22. Conversely, platforms also face potential sanctions in cases of non-compliance. But, what happens if a user challenges the wrongful moderation (Klos, 2020) of content based on the requests of a trusted flagger? Will the online platform remain fully liable as the actor that ultimately removes content? Or is that liability somewhat tempered by the involvement of a trusted flagger? Without users being able to identify trusted flaggers as the source of notices against their content, they are unable to place the liability – at least partly – with that trusted flagger, if necessary. Of course, there can be good reasons to maintain the anonymity of the flagging entity: flagging content for removal can be experienced as an antagonistic action (Myers West, 2018, p. 4373) which may lead to adverse actions against the trusted flagging organisations. However, since trusted flaggers are publicly certified and their identity is accessible via the Commission's database, it makes sense to include their identity in statements of reasons, in order to allow scrutiny of a publicly certified organisation.

Section III: How can we better safeguard freedom of expression in - and ensure effective operationalisation of article 22 DSA?

Despite the implementation of Article 22 being in its infancy, already some recommendations can be made to how it is operationalised and to ensure adequate safeguards for freedom of expression. An initial suggestion was made in the previous paragraph: include the identity of the trusted flagger on the statement of reasons submitted to the user under Article 17 DSA when their content is restricted, to enhance transparency and allow scrutiny of trusted flagger practices. This section proposes additional possible steps that address some deficiencies in Article 22 in the context of Article 22(8) DSA, in which the EU regulator is empowered to further specify how Article 22(2), (6) and (7) must be applied. This article offers three specific recommendations: the first relates to guidance on the application of Arti-

cle 22(2), the second relates to funding and independence, the third relates to the operationalisation of Article 16 for certified trusted flaggers under Article 22.

(i) Harmonising selection criteria and procedure

Firstly, on selecting trusted flaggers and the requirements for certification. The Board of DSCs – a board made up of all DSCs and the European Commission established in Article 61 DSA – could create standard procedures on the appointment of trusted flaggers, particularly on the timeframes and appealability of such decisions, to equally facilitate organisations across the EU. The application procedure could be harmonised as well. Currently, it is fragmented: in some jurisdictions, for example Finland, organisations seeking certification can get into direct contact with the DSC to submit their application, whereas in other jurisdictions, such as Denmark and Austria, they must use a form. A standardised form that ensures that similar documentation is required from organisations across the EU would be a logical next step. Accompanying the development of such a form, the Board should release guidelines on the interpretation of requirements such as 'independence' and 'expertise' to apply a similar standard of proof across the EU. Independence as a requirement pertains to whether trusted flaggers can be influenced by online platforms in their operation – such would forego their neutral position in content moderation. However, in practice, some trusted flaggers are funded partly by online platforms. The best-known example is the INHOPE network, which is partly funded by big tech companies such as Meta and TikTok. The degree to which such funding is allowed under Article 22(2) is contentious: funding by a tech company could be read as a contraindication of independence. However, safeguarded by a strong organisational structure, operational independence can still be guaranteed. For example, the Austrian DSC has certified two entities that are partly funded by online platforms, because the funding was not deemed significant enough to harm independence (KommAustria KOA 16.400/24-017; KOA 16.400/24-013). To facilitate trusted flaggers in combatting illegal content, DSCs must take into account funding opportunities for their work, and consider that funding does not necessarily interfere with the principle of independence. The Board should therefore provide guidance on how organisational structures for trusted flaggers can ensure independence while allowing for trusted flaggers to be sufficiently funded, also by industry partners. Finally, the Board should introduce a 'backstop' in the granting of trusted flagger statuses, in cases when selected organisations could form a challenge to the rule of law, for example, organisations that are expected to flag content on the basis of national law that is not compliant with EU law. An example is a trusted flagger certified in Hungary to flag content contravening Hungarian anti-LGBT legislation: illegal content in that member state

would not be illegal in the rest of the EU. This 'backstop' could also ensure safeguarding of the 'objectivity requirement' of Article 22(2)(c) discussed above. Using the backstop, the Board has a tool to ensure that trusted flaggers are not overly exercising their privilege at the expense of a balanced approach to freedom of expression. This could be by either limiting the number of trusted flaggers altogether, or by (re)assigning trusted flagger statuses to entities that represent an underrepresented interest, for example in the unlikely scenario that only IP organisations would be selected as trusted flaggers. Such standards can be better safeguarded if the Board has a backstop to revoke trusted flagger status, for example, by request of a majority of the DSCs, in lieu of the accrediting DSC being solely responsible.

(ii) Funding

A second improvement that could be made to the Article 22 DSA framework is funding. As indicated under III.i, funding can be a contentious issue due to the independence requirements trusted flaggers have under Article 22. However, in order to operate, trusted flaggers require resources. A prime candidate to supply funding should be the providers of online platforms – ultimately beneficiaries of their services since trusted flaggers contribute to the detection of illegal content. Although providers of online platforms fund some trusted flaggers already, it is worthwhile investigating whether certified trusted flaggers could benefit from a general support fund managed by the European Commission. A model for this is already present in the DSA: Very Large Online Platforms and Very Large Online Search Engines provide a supervisory fee to fund the Commission's oversight role under Article 43 DSA. In addition to providing better resources to trusted flaggers to combat illegal content, restructuring funding can form an incentive for providers of online platforms to better engage with trusted flaggers, since they are the ones funding their operation. To operationalise this structure, the Commission could set up a fund requiring VLOPs to contribute and divide it equally across certified trusted flaggers. This also creates an incentive for entities to apply for certification with their DSC. Funding restructuring could therefore improve the power balance between the trusted flagger and the platform.

(iii) Creating a harmonised API for trusted flaggers

Thirdly, the DSA leaves platforms discretion to choose the mechanisms for trusted flaggers in line with Article 16 DSA. For example, the Dutch DSC suggests in their DSA guidance that online platforms can facilitate trusted flaggers' mission by appointing staff to deal with notices, or create a designated web interface. The latter

approach reflects how some platforms have suggested they comply with Article 22 (*Conference's notes - Getting Ready for the DSA*, 2024). In principle, requirements for submitting notices as a trusted flagger are harmonised under the requirements of Article 16 DSA. However, in practice the mechanisms created to comply with Article 16 can differ significantly according to the platform (e.g. Holznagel, 2024). Earlier research in the context of the *NetzDG* has shown that the design of flagging mechanisms can affect moderation practices (e.g. Heldt, 2019, p. 12). A standardised application programming interface ('API') across platforms for trusted flaggers could mitigate some of the fundamental rights concerns raised above relating to the privilege with which those entities flag. It has been suggested that standardising APIs could also be used to ensure DSA compliance (Goanta et al., 2022). In creating an API for flagging, it is possible to standardise requirements for flagging content across platforms beyond Article 16 DSA (Husovec, 2024, p. 461) whilst simultaneously ensuring compliance of those flags with requirements necessary for limiting freedom of expression. A standardised API can further be used for transparency reporting, since data that needs to be reported on can simply be exported from the API. In cases of state entities acting as trusted flaggers, the API can require additional input that resembles the requirements needed for interference with freedom of expression: legitimate aim, prescribed by law and necessity in a democratic society (art. 52(1) CFR & 10(2) ECHR). Given that legitimate aims are exhaustively listed under art. 10(2) ECHR (Gerards, 2023, p. 100), it should be possible for state entities to select such aims through this portal. Prescription by law can also be a required part of the notice, both for state and non-state entities. This ensures both compliance with the requirement to only flag illegal content under Article 16 DSA and the requirement of basis in law under Article 10(2) ECHR, also allowing users to assess on the basis of what national law their content is being restricted.

The intended remedy the trusted flagger seeks can also be a requirement in this standardised API. Interferences with freedom of expression must be necessary in a democratic society, which is a requirement of proportionality (*Animal Defenders International v the United Kingdom,* 48876/08, (2013), ECtHR, para 100). Such proportionality can only be assessed if the sought remedy is indicated. Content moderation remedies are diverse, and go far beyond simple 'keep up' or 'take down'. Demonetisation, demoting, and labelling are all effective measures to target unwanted content (Goldman, 2021). In some instances, it can be more effective to label content than to take it down. In contentious cases, where the illegality is not a certainty, this can ensure that freedom of expression is respected while simultaneously ensuring that harmful content is combatted (Morrow et al., 2022). The propor-

tionality assessment of the trusted flagger should also be shared with the user whose content is affected, ensuring they are informed and able to seek redress.

The harmonisation of flagging portals and inclusion of fundamental rights standards can be difficult to realise from a practical perspective, but the DSA provides a legal basis for the Commission and the Board to promote the development of a harmonised flagging API (Article 44(1)(c)). It would require providers of online platforms to agree to be part of the same infrastructure. Such would interfere with their freedom to run a business, and require significant resources. From the perspective of the trusted flagger, elaborate requirements in flagging APIs can also be difficult and inhibit expedience. Trusted flaggers have submitted 300.000 statements of reasons over the past 30 days (per 19 November 2024).[1]Combating illegal content online is a matter of expediency and scalability. It has been proposed that this is why the direct application of fundamental rights norms to content moderation is not practical – fundamental rights assessments do not scale to the enormous volume content moderation requires (Douek, 2020; Sander, 2020). A trusted flagger portal requiring extensive documentation would, therefore, only intensify the scalability critiques raised in the context of trusted flaggers. To remedy this, it is possible to limit the extensive documentation requirements in the harmonised API to government entities certified as trusted flaggers under Article 22, as that is where the most significant risk for fundamental rights lies. Dutch oversight authorities have reported submitting up to 500 notices per year (*Kamerstukken* 2022-2023, nr. 1599), suggesting that elaborate documentation, at least for government entities, is achievable.

Conclusion

This contribution has identified existing and newly arisen risks of the trusted flagger framework, and proposed measures to optimise trusted flagger involvement in content moderation under Article 22 DSA whilst introducing further safeguards around freedom of expression of internet users. Relying on trusted flaggers is a proven method for online platforms to help legitimise content moderation practices. Previously, trusted flagger frameworks were bilateral partnerships between governmental- and non-governmental organisations and platforms. Such arrangements, while ultimately regarded as beneficial to the content moderation process, suffer from transparency, rule of law and freedom of expression concerns – the latter two particularly in the case of involvement of state entities. Article 22 DSA for-

1. The DSA transparency database does not allow for making a distinction as to whether those include all trusted flaggers or only Article 22 certified trusted flaggers.

malises the trusted flagger framework, creating certified trusted flaggers for the EU. The framework of Article 22 leads to more transparency and legitimacy of trusted flaggers, and helps trusted flagger organisations in the operationalisation of their practices. However, Article 22 has not fully addressed concerns for over-removal, freedom of expression, and rule of law. Additionally, different applications of Article 22 across member states can lead to a diffuse trusted flagger landscape in the EU, exacerbating freedom of expression concerns. This article proposes ways in which Article 22(8) can be used to harmonise the application of Article 22 throughout the EU. Further, the European Commission should investigate whether it is possible to create a uniform means for flagging content for trusted flaggers. In order to fully channel the potential of trusted flaggers, the Commission could set up a funding structure funded by VLOPs to ensure that trusted flaggers can contribute more effectively while simultaneously creating a compliance incentive. Using trusted flaggers is a proven way to expeditiously target illegal content based on the expertise of civil society and government entities; this expediency of removing content may not come at the cost of disregarding safeguards around potential interferences with the right to freedom of expression, especially when practical and easily attainable solutions are available.

ACKNOWLEDGEMENTS

The author is grateful to the participants in the DSA and Platform Regulation Conference 2024 for their feedback on a first presentation of these findings, Dr. Taylor Annabell for her feedback on the final version, and the Internet Policy Review reviewers/editors Valère Ndior, Julian Rossi, Anne Bellon, and Frédéric Dubois for their helpful comments.

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Appelman, N., & Leerssen, P. (2022). On trusted flaggers. *Yale Journal of Law & Technology*, *24*, 452–475.

Bloch-Wehba, H. (2019). Global platform governance: Private power in the shadow of the state. *SMU Law Review*, *72*(1), 28–79.

Caplan, R. (2018). *Content or context moderation? Artisanal, community-reliant, and industrial approaches*. Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2018/1

1/DS_Content_or_Context_Moderation.pdf

Casarosa, F. (2021). When the algorithm Is not fully reliable: The Collaboration between technology and humans in the fight against hate speech. In H.-W. Micklitz, O. Pollicino, A. Reichman, A. Simoncini, G. Sartor, & G. De Gregorio (Eds.), *Constitutional challenges in the algorithmic society* (1st ed., pp. 298–314). Cambridge University Press. https://doi.org/10.1017/9781108914857.016

Castets-Renardt, C. (2020). Algorithmic content moderation on social media in EU law: Illusion of perfect enforcement. *University of Illinois Journal of Law Technology & Policy*, *2020*(2), 283–324.

Central European Digital Media Observatory. (2024). *Roundtable on the Digital Services Act*. Centre for law, technology and digitisation. https://www.prf.cuni.cz/sites/default/files/soubory/2023-10/kulaty_stul_EN.pdf

Chang, B. (2018). From internet referral units to international agreements: Censorship of the Internet by the UK and EU. *Columbia Human Rights Law Review*, *49*(2), 113–212.

Citron, D. K. (2018). Extremist speech, compelled conformity, and censorship creep. *Notre Dame Law Review*, *93*(3), 1035–1072.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dang, S. (2022). Twitter dissolves Trust and Safety Council. In *Reuters*. https://www.reuters.com/technology/twitter-dissolves-trust-safety-council-2022-12-13/

De Gregorio, G. (2022). *Digital constitutionalism in Europe: Reframing rights and powers in the algorithmic society* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009071215

Douek, E. (2020). The limits of international law in content moderation. *SSRN Electronic Journal*, 37–76. https://doi.org/10.2139/ssrn.3709566

Dvoskin, B. (2022). Representation without elections: Civil society participation as a remedy for the democratic deficits of online speech governance. *Villanova Law Review*, *67*(3), 447–508.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, *16*(1), 18–84.

Eghbariah, R., & Metwally, A. (2021). Informal governance: Internet referral units and the rise of state interpretation of terms of service. *Yale Journal of Law & Technology*, *23*, 542–617.

Eifert, M., Von Landenberg-Roberg, M., Theß, S., & Wienfort, N. (2020). *Netzwerkdurchsetzungsgesetz in der Bewährung: Juristische Evaluation und Optimierungspotenzial [The German Network Enforcement Act put to the test: Legal evaluation and optimisation potential]*. Nomos.

Elkin-Koren, N., Gregorio, G. D., & Perer, M. (2022). Social media as contractual networks: A bottom up check on content moderation. *Iowa Law Review*, *107*(3), 987–1050.

Flyverbom, M. (2016). Transparency: Mediation and the management of visibilities. *International Journal of Communication*, *10*, 110–122.

Gerards, J. (Ed.). (2023). *Fundamental rights: The European and international dimension* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009255721

Gill, K. (2021). Regulating platform's invisible hand: Content moderation policies and processes. *Wake Forest Journal of Business and Intellectual Property Law*, *21*(2), 171–212.

Goanta, C., Bertaglia, T., & Iamnitchi, A. (2022). *The case for a legal compliance API for the enforcement of the EU's Digital Services Act on social media platforms* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2205.06666

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, *28*(1), 1–60. https://doi.org/10.36645/mtlr.28.1.content

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Griffin, R. (2023a). From brand safety to suitability: Advertisers in platform governance. *Internet Policy Review*, *12*(3). https://doi.org/10.14763/2023.3.1716

Griffin, R. (2023b). Public and private power in social media governance: Multistakeholderism, the rule of law and democratic accountability. *Transnational Legal Theory*, *14*(1), 46–89. https://doi.org/10.1080/20414005.2023.2203538

Haupt, C. (2024). The horizontal effect of fundamental rights. In G. De Gregorio, O. Pollicino, & P. Valcke (Eds.), *The Oxford handbook of digital constitutionalism* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198877820.013.38

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1398

Holznagel, D. (2021). *Safeguarding adequate rights enforcement through the Digital Services Act (DSA)*. HateAid. https://hateaid.org/wp-content/uploads/2022/05/2021-09-08-DSA-How-to-improve-enforcement-through-the-DSA.docx.pdf

Holznagel, D. (2024). How to apply the notice and action requirements under Art. 16(6) sentence 1 DSA – which action actually? *Computer Law Review International*, *25*(6), 172–179. https://doi.org/10.9785/cri-2024-250604

Husovec, M. (2024). *Principles of the Digital Services Act*. Oxford University Press.

Kaesling, K. (2022). Vertrauen als Topos der Regulierung vertrauenswürdiger Hinweisgeber im Digital Services Act [Trust as a topos for the regulation of trusted whistleblowers in the Digital Services Act]. *UFITA*, *86*(2), 328–351. https://doi.org/10.5771/2568-9185-2022-2-328

Kaushal, R., Van De Kerkhof, J., Goanta, C., Spanakis, G., & Iamnitchi, A. (2024). Automated transparency: A legal and empirical analysis of the Digital Services Act transparency database. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1121–1132. https://doi.org/10.1145/3630106.3658960

Kaye, D. (2016). *Report of the Special Rapporteur on the promotion of the right to freedom of opinion and expression* (No. A/HRC/32/38; Issue A/HRC/32/38, p. 23). Human Rights Council. https://primarysources.brillonline.com/browse/human-rights-documents-online/promotion-and-protection-of-all-human-rights-civil-political-economic-social-and-cultural-rights-including-the-right-to-development;hrdhrd99702016149

Kaye, D. (2019). *Speech police: The global struggle to govern the internet*. Columbia Global Reports. https://doi.org/10.2307/j.ctv1fx4h8v

Kerkhof, J. (2024). Jawboning content moderation from a European perspective. In C. Oirsouw, J. Poorter, I. Leijten, G. Schyff, M. Stremler, & M. Visser (Eds.), *Constitutional law in the digital era* (1st ed., Vol. 5). T.M.C. Asser Press.

Kerkhof, J. (2025). Musk, techbrocracy, and free speech. In *Verfassungsblog*. https://verfassungsblog.de/musk-techbrocracy-and-free-speech/

Kerkhof, J. (forthcoming). Constitutional aspects of trusted flaggers in the Netherlands. In J. Dijck, K. Es, A. Helmond, & F. Vlist (Eds.), *Governing the digital society: Platforms, artificial intelligence, and public values*. Amsterdam University Press.

Klos, M. (2020). 'Wrongful moderation': Aansprakelijkheid van internetplatforms voor het beperken van de vrijheid van meningsuiting van gebruikers ['Wrongful moderation': Liability of internet platforms for limiting users' freedom of expression]. *Nederlands Juristenblad*, *95*(43), 3314–3332.

Kosta, E., & Brewczyńska, M. (2019). Government access to user data: Towards more meaningful transparency reports. In R.-M. Ballardini & O. Pitkänen (Eds.), *Regulating industrial internet through IPR, data protection and competition law* (pp. 253–274). Kluwer Law International.

Leerssen, P. (2020). The soap box as a black box: Regulating transparency in social media recommender systems. *SSRN Electronic Journal*, *11*(2). https://doi.org/10.2139/ssrn.3544009

Leerssen, P. (2024). Outside the black box: From algorithmic transparency to platform observability in the Digital Services Act. *Weizenbaum Journal of the Digital Society*, *4*(2). https://doi.org/10.34669/WI.WJDS/4.2.3

Lemley, M. (2021). The splinternet. *Duke Law Journal*, *70*(6), 1397–1428.

Mauritz, F. (2024, April 8). *To define is just to define*. Verfassungsblog. https://verfassungsblog.de/to-define-is-just-to-define/

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*(10), 1365–1386. https://doi.org/10.1002/asi.24637

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Orlando-Salling, J., & Bartolo, L. (2023). The Digital Services Act as seen from the European periphery. In *DSA Observatory*. https://dsa-observatory.eu/2023/10/05/the-digital-services-act-as-seen-from-the-european-periphery/

Quintais, J. P., De Gregorio, G., & Magalhães, J. C. (2023). How platforms govern users' copyright-protected content: Exploring the power of private ordering and its implications. *Computer Law & Security Review*, *48*, 105792. https://doi.org/10.1016/j.clsr.2023.105792

Raue, B. (2023). Artikel 22 – Vertrauenswürdige Hinweisgeber [Article 22. – Trusted flaggers]. In F. Hofmann, B. Raue, M. Dregelies, K. Grisse, F. Hofmann, & K. Kaesling (Eds.), *Digital Services Act: Gesetz über digitale Dienste [Digital Services Act: Law on digital services]* (1. Auflage).

Régulation du Numérique. (2024, February 8). *Conference's notes—Getting ready for the DSA: The role of trusted flaggers for a better moderation of platforms*. https://regulation-tech.cnam.fr/summary-arcom-event-on-january-26th-2024-getting-ready-for-the-dsa-the-role-of-trusted-flaggers-for-a-better-moderation-of-platforms/

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

Sander, B. (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.

Scheppele, K. L. (2023). The treaties without a guardian: The European commission and the rule of law. *Columbia Journal of European Law*, *29*(2), 93–185.

Schneider, J. (2024, October 16). *Woher kommt der Wirbel um 'Trusted Flagger'? [Why all the fuss about 'trusted flaggers'?]*. ZDFheute. https://www.zdf.de/nachrichten/politik/deutschland/trusted-flagger-bundesnetzagentur-zensurvorwurf-100.html

Schwemer, S. F. (2019). Trusted notifiers and the privatization of online enforcement. *Computer Law & Security Review*, *35*(6). https://doi.org/10.1016/j.clsr.2019.105339

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *The Academy of Management Review*, *20*(3), 571–610. https://doi.org/10.2307/258788

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives* (1st ed.). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781108666428/type/book

Trujillo, A., Fagni, T., & Cresci, S. (2023). The DSA transparency database: Auditing self-reported moderation actions by social media. *Proceedings of the 28th 2025 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'25)*. https://doi.org/10.48550/ARXIV.2312.10269

Urban, J. M., & Karaganis, J. (2016). Notice and takedown in everyday practice. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2755628

Urman, A., & Makhortykh, M. (2023). How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. *Telecommunications Policy*, *47*(3). https://doi.org/10.1016/j.telpol.2022.102477

Veil, W. (2024). Verpestet ist ein ganzes Land … (Teil I) [A whole country is polluted … (Part I)]. *CRonline*. https://www.cr-online.de/blog/2024/10/16/verpestest-ist-ein-ganzes-land-teil-i/

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies