Cuadernos de RES PUBLICA en derecho y criminología
Nº en prensa ISSN: 2990-0697 DOI: 10.46661/respublica.12588.

Regulating online content. The evolving role of Trusted Flaggers and its implications for freedom of expression

Regulación de los contenidos en línea. La evolución del papel del alertador fiable y sus implicaciones para la libertad de expresión

Cristina Ortega Giménez
Universidad Miguel Hernández. Elche (España)
c.ortega@umh.es
ORCID: 0000-0001-5422-8463

Resumen

The Digital Services Act (DSA) has introduced the figure of trusted flaggers, entities with expertise in detecting illegal online content whose notifications must be given priority attention by digital platforms due to their degree of trustworthiness (art. 22 DSA). This function gives entities a certain capacity to control public discourse. In this research, we delve into the legal and constitutional nature of these private powers and analyse whether their powers constitute a mechanism for monitoring harmful content, which contributes to the maintenance of peaceful coexistence in a healthier digital environment, or whether the delegation of these responsibilities, under the DSA, entails a form of censorship on freedom of expression, with the dangerous effect it entails for pluralism and democracy.

Palabras clave: Digital Services Act, Content moderation, Constitutional Rights, Freedom of expression, Trusted flaggers, Social Networks.

Abstract

La Ley de Servicios Digitales (DSA) ha introducido la figura de los denunciantes de confianza, entidades con experiencia en la detección de contenidos ilegales en línea cuyas notificaciones deben recibir una atención prioritaria por parte de las plataformas digitales debido a su grado de fiabilidad (art. 22 DSA). Esta función otorga a las entidades una cierta capacidad para controlar el discurso público. En esta investigación, profundizamos en la naturaleza jurídica y constitucional de estos poderes privados y analizamos si sus facultades constituyen un mecanismo de control de contenidos nocivos, que contribuye al mantenimiento de la convivencia pacífica en un entorno digital más saludable, o si la delegación de estas responsabilidades, en virtud de la DSA, supone una forma de censura de la libertad de expresión, con el peligroso efecto que ello conlleva para el pluralismo y la democracia.

Key words: Interoperability: Military forces; Latin America; multi-domain operations; Dominican Republic.

Cómo citar este trabajo: Ortega Giménez, Cristina. *(2026).* Regulating online content. The evolving role of Trusted Flaggers and its implications for freedom of expression. *Cuadernos de RES PUBLICA en derecho y criminología, (en prensa), 01–13.* https://doi.org/10.46661/respublica.12588.

Recepción: 03.09.2025 **Aceptación:** 22.12.2025 **Publicación:** 15.01.2026

1. Introduction

One of the thoughts that has marked my life is Balkin's formulation that the purpose of freedom of expression is to protect and promote democratic culture[1]. This is not a trivial statement: it implies that freedom of expression, beyond its dual dimension (individual and collective), stands as the pillar that allows citizens to participate, on an equal footing, in the processes of construction of meanings and social interactions that define us as individuals.

The beginnings of the Internet seemed to favour the development of freedom of expression in this direction: the neutral and horizontal openness offered by the Net made possible a scenario of pure expression where all discourses competed in an open and unregulated market (Wu, 2003), which in turn meant greater opportunities for those who could not access the conventional media to make themselves heard. This process, it was assumed, would contribute to the strengthening of democratic values, since for the necessary formation of public opinion, the Internet would provide a space for wider and more inclusive debate.

However, over time it has been shown that most of these assumptions proved to be wrong (Vázquez Alonso, 2022), especially the one that proclaimed that the digital environment favoured equal freedom for all users. Of all the reasons explaining why the Internet has become a *liberties' pollution*[2], this research focuses on the consequences of entrusting private actors with the responsibility of ensuring such a democratic culture on the Net (Balkin, 2017, p.2). According to the American jurist, these entities exercise public functions not to safeguard rights such as freedom of expression, but exclusively to implement new forms of control over individuals that help them make their own business models profitable. In this kind of 'surveillance capitalism' (Teruel Lozano, 2023, p.182) where citizens' rights are treated as market products, the protection of freedom of expression determines not only what can and cannot be communicated digitally, but also the nature of the democratic society in which we aspire to live.

In line with these ideas, there is an increasing tendency for public authorities to delegate to digital platforms (such as social networks) the control of online discourse (Teruel Lozano, 2023), a practice that Vázquez Alonso (2022, p.115) calls 'vicarious censorship': cases in which governments attribute to private entities powers that are their own. A highly interesting example is the figure of trusted flaggers, entities specialised in the detection and notification of illegal online content whose influence on the moderation and development of public discourse is dangerously significant.

Beyond Balkin's assertion that private companies have not proven to be reliable stewards of the values of freedom of expression in the 21st century (2017, p. 3), this research poses several questions: are private powers legitimised to protect fundamental rights in the digital space? In particular, what are the challenges to the exercise of freedom of expression posed by the functions that these organisations perform? Finally, are there ways in which, in the specific case of trusted flaggers, the control of public discourse on the Internet can be an instrument with full constitutional guarantees?.

1.1. The horizontal effect of fundamental rights on private relationships

In today's democracies, fundamental rights are not exercised exclusively in the state-society dichotomy (Cruz Villalón, 1989; Sarazá Jimena, 2008). Private actors, especially technological corporations such as intermediary service providers, have taken on

[1] Balkin, J. M. (2017). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation, *UC Davis Law Review, Yale Law School, Public Law Research, 615, http://dx.doi.org/10.2139/ssrn.3038939.*

[2] Pérez Luño (2024, p. 99) alludes to the term from English social theory to explain that the extensive use of the Internet has produced a degradation of fundamental rights.

a decisive role in shaping the digital public space, controlling and conditioning access to information, freedom of expression and the right to privacy. This scenario has reopened the debate on the horizontal application of fundamental rights, which German doctrine conceptualised as *Drittwirkung*, according to which, although fundamental rights are originally conceived to limit the power of the State, they can also have effects in relation to third parties.

Constitutional dogma has been discussing for more than sixty years whether fundamental rights are predicated only against the State or whether they can also be enforced in the private sphere, without there being any unanimity in this regard to date (Naranjo de la Cruz, 2000; Bilbao Ubillos, 2017).

The problem stems from our Constitution which, unlike the Portuguese Constitution, whose Article 18 provides that the rights and freedoms recognised therein are directly applicable to public and private entities, our 'Carta Magna' does not expressly address this issue. It is true that Article 53.1 Spanish Constitution establishes that the rights and freedoms recognised in Chapter II of Title I are binding on all public authorities. However, it does not seem that it can be deduced from this constitutional precept that these rights are not directly binding on individuals in any case, as such a conclusion would be incompatible with art. 9.1 EC, which states that both public and private authorities are subject to the Spanish Constitution (Beladíez Rojo, 2017, p.78).

The majority of the doctrine seems to be convinced of a horizontal expansion of fundamental rights, and constitutional jurisprudence has pronounced in the same sense. From their first judgments, the Constitutional Court of Spain recognised that when fundamental rights are exercised between private individuals, their content and exercise are also subject to specific limits. STC 18/1984, of 7 February, which constitutes the *leading case* in this matter, established in its 6th FJ that public authorities' subjection to the Constitution *translates into a positive duty to give effect to such rights in terms of their validity in social life.* This was also ratified later in STC 177/1988, of 10 October, (FJ 4º): *In a social State governed by the rule of law, it cannot be maintained in general that the holder of such rights is not the holder of those rights in social life. This is why, this Court has recognised that private acts can infringe fundamental rights.*

Therefore, the content of fundamental rights extends to relations between citizens and can be claimed against this type of subject.

The extension of legal protection into the private sphere aims to establish a minimum standard of protection in all spheres to avoid the existence of areas of impunity where rights are not respected. It also aims to adapt the recognition of these rights to new social realities by promoting a constitutional culture that must go beyond the state level. Above all, the *Drittwirkung* allows for the imposition of obligations on private actors with representative power in certain areas, i.e. in a vertical or dominant relationship with the user, as is the case on the Internet. In this sense, technology corporations constitute 'an ideological power characterized by the gift of speech, the primary instrument of domination' (Gutiérrez Gutiérrez, 1999, p.204), but 'formally private, with forms of coercion analogous to those of public powers' (Bilbao Ubillos, 2017, p.51). The decisions of these Internet giants, which are attributed broad powers of self-protection, are as imperative and immediately enforceable as those adopted by an administrative body.

Thus, the objective is that digital corporations (such as social networks) do not act with their backs turned to the restrictions imposed by the Spanish Constitution to protect the dignity and essential freedoms of the general population.

In the current digital context, the application of this doctrine raises a key question: to what extent must digital platforms, as hegemonic private powers, respect and promote the exercise of fundamental rights?

1.2. Scope and effectiveness of *Drittwirkung* on the Internet: its application in social networks

To answer the previous question, we must recognize that we are witnessing a deep crisis of the public-private dichotomy, where the

boundaries between the two spheres have become blurred (Bilbao Ubillos, 2017, p.51-52). Thus, public power tends to become privatized, while private power increasingly assumes public connotations, giving rise to a progressive 'intersection' between both (Gutiérrez Gutiérrez, 1999, p.205). Currently, digital platforms represent this intersection, as they have become hybrid (public-private) forums.

According to Vázquez Alonso (2022, p.122), these companies have created discussion spaces that are unequivocally public in nature, aligning with the basic principles of our democratic system. In this way, through their technology, they condition the public sphere of social communication, which requires states to intervene and regulate them by imposing limits on their content control policies to protect democratic pluralism and the freedoms of expression and information.

However, these digital companies also retain a private character derived from the freedom of enterprise, which gives them the power to decide how to operate according to their commercial interests. Therefore, they cannot be denied a certain autonomy to manage their content and set rules of behaviour for their members without this necessarily constituting a form of censorship. This power is manifested, in particular, through the Terms and Conditions of Service, which constitute a contract of adhesion that users accept as a condition of access to their services.

This public-private duality of social networks implies two possible applications of *Drittwirkung* in their operations:

a) Direct effectiveness: fundamental rights are applied directly in private legal relations, especially if an asymmetry of power between the participating subjects can be demonstrated. One doctrinal sector argues that digital platforms, due to the aforementioned public dimension they possess and, therefore, their influence on the functioning of democracy, should be treated as public services (Peña Jiménez, 2021, p.294). This interpretation views social networks as quasi-state actors, which would oblige them to respect the fundamental rights of users with a standard as high as that required of the State. This formulation is in line with the thinking of Sunstein (2017), who argues for a robust application of fundamental rights in the digital sphere. The author is concerned, from a democratic perspective, that citizens are not sufficiently exposed to content and viewpoints beyond those they select based on their own preferences (echo chambers). He proposes an active intervention in platforms by configuring them in such a way that they expose users to a variety of perspectives that encourage public debate and avoid polarisation. Therefore, he proposes government regulation of the Internet as a public good, not in a traditional sense that would completely eliminate the autonomy of private actors, but to implement an architectural structure of social networks designed on the basis of principles that benefit collective deliberation and the quality of democracy. Thus, Sunstein's proposal to regulate digital platforms in order to foster pluralistic debate is in line with the idea that fundamental rights generate automatic obligations and are directly applicable in cyberspace.

b) Indirect effectiveness: fundamental rights must be taken into account when private companies interpreting and applying private law. This means that, although social networks are private entities, their rules and decisions on content moderation must respect constitutional rights such as freedom of expression. This would frame the reasoning of Suzor (2017) who emphasises the need to strike a balance between the autonomy of private actors and the imperative to protect fundamental freedoms. And since he considers constitutionalism as a limitation of powers, he proposes a digital type of constitutionalism that, on the basis of human rights, is capable of limiting the supremacy of online platforms, legitimising their governance processes and improving transparency in the decision-making process on content moderation. In this way, and although in this interpretation the effectiveness of fundamental rights may not fully apply to the private relationships that take place online, there needs to be a harmonisation between the community rules of social networks and constitutional values.

We consider that the reinforced application of the indirect version of the *Drittwirkung* is more appropriate, i.e. fundamental rights do not directly bind social networks as if they were public institutions, but they should serve an essential criteria in interpreting the platforms' rules and decisions. As proposed by Vázquez Alonso (2022), a high degree of harmonisation is required between the content rules of social networks and the existing constitutional understanding of the freedoms of expression and information. This means that social networks should not restrict content more strictly than is established in the Spanish Constitution, nor should they apply rules of behaviour that contradict fundamental rights.

The direct application of the *Drittwirkung* could be extended in such a way as to conceive of social networks as quasi-state actors, which would mean equating the regulation of their terms of service with the exercise of public functions. Such a view would create tensions with the principle of entrepreneurial autonomy and blur the distinction between state and private actors; a circumstance that would even lead to the former using the latter in a devious way to curtail the freedom of the population and impose 'one-dimensional thinking' (Herbert Marcuse, 1964).

Despite this, we agree with Sunstein (2017) on the power of social networks to condition public debate, to the extent that they can determine which voices participate and which are silenced through the echo chambers or bubble filters that they algorithmically design to intellectually isolate their users. Networks, as we have already said, are not mere private forums, but fundamental spaces for the exercise of rights such as freedom of expression and access to information, which is why they must respect certain principles that guarantee pluralism and due process in the moderation of content. But the purpose should not be the reconfiguration of private law, but rather to establish guarantees to protect areas of human freedom in environments where they are threatened (Hesse, 2001).

If we apply the *Drittwirkung* indirectly, we make it clear that the constitutional framework must transcend the State-Society dichotomy, with fundamental rights being an interpretative criterion in the activity of social networks, but without replacing the judiciary as the ultimate guarantor of those rights. What is certain is that this approach strengthens the role of the judiciary because it prevents private companies alone from regulating and defining the limits of freedom of expression in the digital environment, thus preventing them from becoming parallel judicial powers that operate without constitutional guarantees.

Finally, in response to the question we posed at the beginning of this paper -are private powers legitimised to protect fundamental rights in the digital space?- the answer must be affirmative, especially when they occupy hegemonic positions from which they not only favour but also influence the exercise of fundamental rights such as freedom of expression. However, this does not imply that they should arrogate to themselves the same obligations as the State, nor that their role should be so decisive that governments should delegate to them the responsibility of facing the challenges posed by the exercise of freedom of communication in the virtual world.

In the first instance, the task of establishing a regulatory framework that balances the fundamental rights at stake falls to the public authorities. It is then up to service providers, such as social networks, to assume their share of responsibility for protecting freedoms. And, ultimately, it should be the judiciary who retains the final word on possible infringements of rights in the digital sphere, since judicial proceedings offer essential guarantees for better safeguarding rights.

2. Trusted flaggers under the Digital Services Act: the new supervisors of the digital space

One of the most recent and significant manifestations of the intervention of private authorities in the protection of fundamental rights in the digital environment is the role of *trusted flaggers*. Introduced by Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services (*Digital Services Act* or DSA), these are private bodies with quasi-public functions, as they have specific

knowledge and skills to detect, identify and report illegal content.

To be granted alerter status, Art. 22 DSA requires, in addition to having sufficient expertise in the field, not to be dependent on any online platform provider and to carry out its activities with the aim of sending notifications in a diligent, accurate and objective way. It is sufficient to note the enormous privilege these entities have in monitoring online speech, which may affect freedom of expression, especially in cases where it is uncertain whether the content constitutes an offence. We are referring to 'dissenting speech': those offensive or harmful expressions that from a moral point of view may be objectionable but do not constitute a crime.

This supervisory power is even greater if we take into account that platforms must prioritise notifications sent to them by these entities through the notice and action mechanisms (Art. 16 DSA), and must resolve them without undue delay.

2.1. Conception and historical background

The existence of bodies to which special trust is given and, consequently, preferential treatment in the reporting of illegal content has always been present in the minds of data hosting platforms (Flaquer Riutort, 2022, p.3). Collaboration with these figures made it possible to improve content moderation processes, since trusted flaggers quickly and effectively pointed out illegal content, which helped social networks demonstrate that they were taking active measures to comply with Internet content regulation and thus avoid possible sanctions. Their work also reduced the number of false or malicious reports that overloaded the platforms' workload and optimised internal processes for detecting harmful content. However, behind these laudable objectives, Cetina Presuel (2024, p.253) explains that there was also 'the [platforms'] appetite for self-governance and their allergy to state regulation and, of course, a self-preservation interest that led them to seek to maintain their exemption from liability in order to sustain their business model'.

The historical antecedents of *trusted flaggers* can be found in US law, specifically in Section 230 of the *Communications Decency Act* and Section 512 of the *Digital Millennium Copyright Act*, the paradigmatic example being YouTube's Content ID[3] . With the adoption of Directive 2000/31/EC on Electronic Commerce, the European Union transposed the same scheme of limited liability and MNAs into European law, which eventually led to the emergence of trusted alerters through mechanisms such as Europol's Internet Referral Unit (EU-IRU)[4].

The European Police Office played an active role in flagging terrorist or illegal content for review, and in most cases, it was removed by the platforms without being declared illegal by a competent judicial authority (Cetina Presuel, 2021, p. 531). This practice entailed the 'adjudication' of the fundamental right to freedom of expression to private companies, with the approval of the public authorities (Cetina Presuel, 2024, p.260).

Beyond the example of Europol as a reliable whistleblower (given that the police force enjoys legitimacy derived from the legal mandate granted by the Member States), is especially relevant the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions of 28 September 2017, which specifically addressed the fight against illegal content online[5], which recognised that illegal content would be removed more quickly and reliably if digital intermediaries implemented mechanisms to provide a seamless

[3] For more information: https://support.google.com/youtube/answer/2797370?hl=en.

[4] For more information: https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru.

[5] Document available at https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52017DC0555&from=SV.

communication channel for reporters better positioned to flag potentially illegal material on their websites. This was one of the first European documents to introduce the concept of 'trusted reporters' (Flaquer Riutort, 2022). However, it was the subsequent Commission Recommendation (EU) 2018/334 of March 2018, on measures to effectively combat illegal content online[6], that incorporated an explicit definition of 'trusted flagger': a *natural or legal person that a data hosting service provider considers to have particular competences and responsibilities for the purpose of combating illegal content* online.

Finally, the Digital Services Act (2022) formally defines the figure of these bodies in its Art. 22; and develop their characteristics in several recitals that will be analysed below. The aforementioned Recommendation and the new Regulation share the need to establish an alert system managed by external actors to prevent the dissemination of unlawful, undesirable and harmful content on social networks and/or to ensure its removal in a fast, accurate and balanced way. Likewise, the European Union understands that if platforms rely on experienced actors, the risk of arbitrary notifications is reduced, and a safe online environment that respects fundamental rights is built.

However, it is the DSA that legally enshrines the figure of trusted flaggers, codifying it for the first time in a strong regulatory measure, such as a rule directly applicable to all EU countries.

2.2. Constitutional characteristics and problems

Art. 22 and recitals (61) and (62) of the DSA set out the characteristics of these privileged notifiers:

A) The lack of precision in the definition of these bodies, the ambiguity in the criteria for determining who can be considered as a trusted flagger, and the procedure for applicants to demonstrate that they can carry out the assigned functions are the main problems with the characterisation of these figures (Cetina Presuel, 2024). Nor does the DSA specify the average time platforms have to process notifications, or whether these will come from monitoring systems with identical standards, applied to all entities that exercise the role of reliable communicators, or whether trusted flaggers will use their own systems for detecting illegal content, which is likely to lead to numerous discrepancies between them. All these issues entail a violation of citizens' right to be properly informed (Art. 20 Spanish Constitution) about the designation and functioning of a figure that, in the end, will act as a content filter, assessing whether a discourse is allowed (or not) in the digital space, with the consequences that this decision entails for the free formation of public opinion. The right to information not only implies access to truthful and plural information about the ecosystem in which we interact, but also the possibility of knowing (and questioning) actions that affect public debate, such as content moderation, which is one of the main tools for controlling discourse (Cotino Hueso, 2023). Similarly, this lack of transparency in the designation of alerters entails a breach of the principle of legal certainty, as the wide margin of appreciation that the DSA confers on countries promotes arbitrariness in their selection, assessment of their capabilities and definition of their functions.

B) Trusted flagger status has to be granted by the digital services coordinator of the Member State where the applicant is established. In the case of Spain, it is the Comisión Nacional de los Mercados y la Competencia (CNMC) that exercises the role of Coordinator. The fact that an administrative body is in charge of appointing the alerters —who, let us not forget, will enjoy a privileged position to monitor and control online speech— raises questions of constitutionality, since freedom of expression is a fundamental right, and all actions which may limit its exercise must be regulated by organic laws. Furthermore, the decisions taken

[6] Document available at https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32018H0334&from=FR.

by the CNMC regarding the selection of trusted flaggers may be biased and favour certain political and economic interests, since the rule is vague when it comes to defining the criteria to be met by applicants (Art. 22.2 DSA), which broadens the margin of appreciation by the digital coordinators. In short, the selection of these entities by this administrative body may give rise to a clear abuse of power on its part. In addition, this situation generates inequalities for applicants and for the rest of the Internet users whose rights may vary depending on how each country applies the DSA, in relation to the selection of the trusted flaggers who, in theory, are supposed to protect them.

C) This status can only be granted to entities and not to qualified individuals. Recital (61) of DSA specifically refers to Europol and the organisations that are part of the INHOPE network[7]. Unlike the aforementioned 2018 Recommendation, which did allow it to be granted to qualified individuals, the DSA reserves this position exclusively for organizations constituted as legal entities. According to González-Orús Charro (2024, p.260), the reason for the exclusion of natural persons may be due to technical reasons: entities have collegiate governing bodies and are in a position to assess high volumes of content more efficiently. In addition, companies are often subject to transparency and oversight processes conducted by competent authorities, whereas natural persons would be harder to monitor and might even be more vulnerable to external pressures. Leaving aside technical issues, this exclusion means limiting the direct participation of citizens in the moderation of digital content, in which they play a leading role, as it is the users who are ultimately affected by content removal policies. This restriction could favour the concentration of power in a few companies with commercial or corporate interests, while depriving certain individuals (e.g. independent activists, journalists, or citizen groups) of the ability to influence the selection of harmful content, which ends up negatively impacting the value of pluralism and the democratic principle.

D) Recital (62) and Art. 22(3) DSA refer to the obligation for trusted flaggers to publish (at least once a year) *easily understandable and detailed* reports on their notifications sent in accordance with this rule. The aim of this measure is to demonstrate that reporters work objectively and independently. However, effective monitoring mechanisms for these reports (a measure on which the DSA is silent) are necessary to prevent them from becoming a mere formality. They must assess the impact of the notifications sent by these bodies, thoroughly ensuring that they do not commit abuses or become censors of freedom of expression in the digital space.

E) The DSA also provides for the revocation of alerting status where there is evidence that organizations have sent a significant number of insufficiently accurate, incorrect or inadequately substantiated notifications (Art. 22(6) and (7) DSA). The European legislator is also unclear about the grounds for such a suspension, which will ultimately be determined by the digital services coordinator. This punitive measure could disproportionately affect trusted flaggers with fewer resources or less capacity to fulfil their function of detecting illegal content. This would be a form of indirect discrimination against smaller entities. Similarly, the decision on revocation is an excessive power that digital coordinators would enjoy and, as such, would need to be subject to due process guarantees. For example, it would be necessary to define precisely what is meant by 'a significant number of notifications', 'inaccurate', 'incorrect' or 'inadequately substantiated' notification and the implications for the organizations concerned (principle of legal certainty). Despite the rule providing that, before revoking such status, the DSA must give the entity an opportunity to respond to the findings of its investigation and its intention to revoke its status. We strongly believe that there is no real appeal or judicial review mechanism

[7] For more information, refer to https://www.inhope.org/EN?locale=es.

to challenge the decision taken by the DSA, and, above all, to protect the rights of alerters against possible administrative arbitrariness.

As we can observe, the European rule on trusted flaggers contains numerous loopholes which perhaps are responsible for the unusual lack of implementation of this figure. Since the DSA finally came into force (February 2024), only a few trusted flaggers have been appointed in the European Union[8], and none in Spain. Goldberger (2024) identifies some problems that may be preventing its implementation: the dissuasive effect of certain bodies becoming `whistleblowers´ or censors of online discourse, especially if they carry out their task in such a way that it is perceived by users as a sign of over-flagging (excessive moderation).

A risk that the DSA does not seem to contemplate, as it does not include any preventive measures in this regard. In addition, some trusted flaggers receive direct funding from the social networks they are supposed to moderate, which may lead to a certain dependency and an obvious conflict of interest that may lead them not to apply for alerter status. As discussed above, the DSA requires full independence from digital corporations in order to confer the status of trusted reporter.

The characteristics of the codification of trusted flaggers and their scarce implementation raise important constitutional questions. Of all those that have already been formulated, we are now interested in addressing, specifically, their impact on freedom of expression, the backbone on which a 'healthy and vibrant' digital space must be built (Balkin, 2017, p.68).

3. Who supervises the supervisor? Trusted flaggers in the face of constitutional protection of freedom of expression

The privatisation of freedom of expression involves shifting responsibility for identifying and requiring the removal of objectionable content to trusted flaggers, which may result in undue restrictions on the fundamental right to express oneself. Firstly, because, as Schwemer (2019, p.9) confirms, private entities can impose ('and be encouraged to impose') restrictions on access to information without being subject to the constitutional limits that apply to the State when it seeks to restrict freedom of expression. Moreover, if these censorship decisions are supported or facilitated by public authorities, it becomes even more difficult to assess whether fundamental rights have been respected. In fact, such restrictions would face strong opposition from the public if they were enshrined in law (Gorwa, 2019, p.13).

We are alluding to an issue already mentioned at the beginning of this research: expressions that shock, disturb or offend the State, or a part of its population, must have their own space reserved and retain their right to participate in the public debate, in order to be known and to be better able to be refuted (Boix Palop, 2016). Otherwise, by justifying public intervention to exclude harmful (but not illegal) speech from the public sphere, we would be defending an 'unbearable legal paternalism' (Alcácer Guirao, 2020, p.262) in which the State views citizens as children needing protection and isolation from certain forms of speech. And not as adults, with sufficient autonomy and critical capacity to discern right from wrong.

This issue connects to a second point to be taken into account: some types of trusted flaggers will have an interest in certain content (harmful speech) not being present on the Internet. In other words, the following dichotomy will arise: trusted flaggers will aim to remove as much content 'as possible' while the social networks only want to remove as much content 'as necessary' (Schwemer, 2019, p.9).

This is due to an economic incentive that encourages trusted flaggers to impose

[8] The list, which is updated as new trusted flaggers are appointed in each EU country, is surprisingly short and can be found at the following link: https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa#The%20list%20of%20DSA%20Trusted%20flaggers.

restrictions on questionable content because it allows them to protect their business model. Delegating moderation power to these figures helps platform operators reduce labour costs while limiting their regulatory liability for extraneous conduct on their service and positioning themselves as 'champions of free speech' (Matias, 2019, p.2). The background to all this is that freedom of expression is commodified, making it secondary to economic interests, something that should be unacceptable from a standpoint based on the defence of fundamental rights (Cetina Presuel, 2024, p.262).

Over-reliance on trusted flaggers is also problematic. Their relationship with the platforms is marked by an 'asymmetry of knowledge' (Schwemer, 2019, p.9), i.e., it is assumed that the alerter knows more (or is better informed) about certain content than the social network, which has to deal in a generalised way with everything posted by users. This asymmetry affects freedom of expression because taking most of the notifications sent by reliable communicators for granted can lead to excessive removal of content. A fact that has already been noted, as Teruel Lozano (2023, p. 216) explains: 'These suggestions for the removal of content, however voluntary they may be, are taken into account in more than 90% of cases.'[9] Undoubtedly, this fact proves the capacity of these entities to 'create pressure and make companies remove content, even if the illegality has not yet been determined by any competent authority' (Cetina Presuel, 2021, p.530).

We agree with this author that these practices do not comply with the constitutional premise which provides that restrictions on rights must be enshrined in law. It is also difficult to understand how freedom of expression can be effectively protected against any possible abuse by social media platforms, as this procedure clearly lacks transparency, due process guarantees and adequate judicial control.

Moreover, if we consider the requirements established by the DSA for acquiring whistleblower status (Art. 22 DSA), it is clear that they favour the representation of influential entities on the Internet. The references in the DSA to Europol as an example of a reliable notifier, as well as the exclusion of natural persons from obtaining this status, discriminate against less powerful stakeholders or those representing the interests of marginalised groups who also want to influence platform governance and contribute to content moderation in their communities.

The due participation of disadvantaged groups, in trusted alerting schemes, would improve the balance of power and reinforce the democratic principle. Also, this fact would be ensuring that not only majority interests are protected in content moderation, but also those of disadvantaged groups who already face significant barriers to accessing public spaces. For example, they could identify problems such as algorithmic discrimination or the marginalisation of digital narratives that may not be prioritised by the most influential alerters. All of this would lead to better protection of fundamental rights, with trusted flaggers acting as guarantors of a secure digital ecosystem rather than as enforcers policing freedom of expression that unsettles certain sections of the population.

4. Conclusions

In this research, we have examined the figure of trusted flaggers, introduced by the Digital Services Act in its Article 22. The analysis of these bodies, focused on the surveillance and detection of illegal content on the Internet, has obliged us to study, firstly, the responsibility of private powers in the protection of fundamental rights. The horizontal application

[9] This is reflected in the Communication from the Commission to the European Parliament, the European Council and the Council in implementation of the European Agenda on Security to combat terrorism and to pave the way for a genuine and effective Security Union (COM/2016/0230 final). It details that Europol has made more than 3200 requests to digital platforms to remove hate speech and terrorist propaganda content, with an effective removal rate of 91%. Document available at https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52016DC0230.

of rights in private-legal relations (*Drittwirkung*) has meant conferring certain powers on digital platforms in the virtual environment, including content moderation. Towards this end, social networks rely on the work of these entities who present several constitutional problems: the lack of precision in their conceptualisation and the obligation for them to be appointed by digital coordinators, based on general and ambiguous requirements, may encourage arbitrariness and bias in their choice, compromising the exercise of the citizens' right to information.

Moreover, their regulatory configuration may violate the right to equality and may generate discrimination towards sectors representing the interests of smaller groups or even towards natural persons, whom the DSA directly excludes from the possibility of obtaining the status of trusted flagger. In this sense, there is an urgent need for a more adequate delimitation of the concept and structure of trusted flaggers, establishing that the status of alerter is granted to a plurality of organisations representing different sectors of society, with the obligatory inclusion of vulnerable groups.

Promoting the value of pluralism and the democratic principle would reduce the risk of these entities acting for the benefit of actors with great power or influence on the Internet. It is also necessary to ensure due process guarantees when they activate the revocation procedure which, under the DSA, is subject to a wide margin of discretion on the part of the digital coordinators, which creates a clear lack of defence for the alerters.

However, the possible infringement of the fundamental right to freedom of expression is the main consequence of the European regulation of this figure. Given that notifications of questionable content are prioritised by the platforms (presumption of veracity) and, in most cases, eliminated, alerters are given exorbitant power to control public discourse.

This practice can lead them to want to remove as much content as possible to fulfil their mandated duties and sustain their business models. This makes them a censor of freedom of expression, especially of those messages that may be controversial ('dissident speech') but which, if they are not forbidden by the legal system in the physical space, they should not be prohibited in the virtual world either.

Likewise, trusted flaggers can become outsourced censorship tools serving public authorities that evade their duty to protect citizens' rights by transferring it to private companies. Instead of establishing a regulatory framework that balances the fundamental rights at stake, public authorities allow certain actors, such as platforms or trusted flaggers, to make content decisions that escape democratic control and ultimately imply a form of privatisation of justice. Because we cannot forget that there is no prior judicial control to evaluate the notifications sent by the trusted flaggers to the digital platforms, nor is there any subsequent control to confirm that the removal of content complied with the constitutional guarantees of the right to freedom of expression.

To curb this power and ensure that communicators submit high-quality notifications, it is crucial to consider some key measures such as restricting automated notifications to cases where false positives compared to human-generated notifications are extremely rare (Schwemer, 2019, p.21). Another measure, which would ensure that these bodies do not become the sole means of speech control, would be to implement independent monitoring and control mechanisms for their functions, for example by obliging them to publish continuous detailed reports on their activities (Art. 22.3 DSA only requires an annual publication).

Trusted flaggers can play a key role as facilitators of a digital space for peaceful and tolerance, provided that their content monitoring is proportionate, verifiable and respectful of the fundamental right to freedom of expression. This work must also be subject to strict oversight and transparency controls. If these safeguards are not implemented, there is a risk that these entities could become dangerous tools, susceptible to generating a 'perverse Brussels effect' (Bradford, 2012; Cetina Presuel, 2024, p.262).

This means that these figures could be exported to dictatorial countries to legitimise their use for non-democratic purposes. Only

the time required for the implementation of the DSA will determine what role these watchdogs will end up playing: will they act as guarantors of a digital ecosystem that fosters peaceful coexistence and protects democratic culture? Or will they act as censors of freedom of expression, undermining democracy itself? Hopefully, and our digital freedom depends on it, the first option will prevail.

Referencias

BALKIN, Jack M. (2017) Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. UC Davis Law Review, Yale Law School, Public Law Research, 615. http://dx.doi.org/10.2139/ssrn.3038939.

BELADÍEZ ROJO, Margarita. (2017) La eficacia de los derechos fundamentales entre particulares. Anuario de la Facultad de Derecho de la Universidad Autónoma de Madrid, 21:75-97. https://repositorio.uam.es/handle/10486/686460.

BILBAO UBILLOS, Juan María. (2017) La consolidación dogmática y jurisprudencial de la Drittwirkung: Una visión de conjunto. AFDUAM, 21:43-74. https://repositorio.uam.es/handle/10486/686459.

BOIX PALOP, Andrés. (2016) La construcción de los límites a la libertad de expresión en las redes sociales. Revista de Estudios Políticos, 173:55-112. https://doi.org/10.18042/cepc/rep.173.02.

BRADFORD, Anu. (2012) The Brussels Effect. Northwestern University Law Review, vol. 107:1. https://ssrn.com/abstract=2770634.

CETINA PRESUEL, Rodrigo. (2024) Alertadores fiables: de su codificación en la DSA a la necesidad de atender sus limitaciones. In: Serrano Maíllo MI (ed) and Corredoira L (ed) Democracia y desinformación: nuevas formas de polarización, discursos de odio y campañas en redes. Dykinson, Madrid, p. 251-266.

COTINO HUESO, Lorenzo. (2023) Menos libertad de expresión en internet: el peligroso endurecimiento del TEDH sobre la responsabilidad de moderación de contenidos y discurso del odio. Digital Law and Innovation Review, 16. https://www.uv.es/cotino/publicaciones/TEDH2023ESCRIBOV2.pdf.

CRUZ VILLALÓN, Pedro. (1989) Formación y evolución de los derechos fundamentales. Revista Española de Derecho Constitucional, 9 (25), 35-62.

FERRAJOLI, Luigi. (2011) Poderes salvajes. La crisis de la democracia constitucional. Editorial Trotta, Madrid. https://doi.org/10.2307/j.ctv31zqgf4.4

FLAQUER RIUTORT, Juan. (2022) El papel de los alertadores fiables en la detección de contenidos ilícitos en la red. Revista Aranzadi de Derecho y Nuevas Tecnologías, 58:1-15.

GOLDBERGER, Inbal. (2024) Europe's Digital Services Act: Where Are All The Trusted Flaggers? Tech Policy. Available via https://www.techpolicy.press/europes-digital-services-act-where-are-all-the-trusted-flaggers/. Accessed 30 March 2025.

GONZÁLEZ-ORÚS CHARRO, Martín José. (2024) El alertador fiable: nueva figura incorporada por el reglamento de servicios digitales para la detección y notificación del contenido ilícito en la red. In: Carbajo Cascón F (ed) and Curto Polo M (ed) Derecho digital y mercado. Tirant lo Blanch, Valencia, p. 235-268.

GORWA, Robert. (2019) The platform governance triangle: conceptualizing the informal regulation of online content. *Internet Policy Review*, 8(2). https://doi.org/10.14763/2019.2.1407

GUTIÉRREZ GUTIÉRREZ, Ignacio. (1999) Criterios de eficacia de los derechos fundamentales en las relaciones entre particulares. Teoría y Realidad Constitucional, 3:193-211. https://doi.org/10.5944/trc.3.1999.6478

HESSE, Konrad. (2001) Manual de Derecho constitucional. Marcial Pons, Madrid.

MARCUSE, Herbert. (1964) El hombre unidimensional. Austral, Barcelona.

MATIAS, J. Nathan. (2019) The Civic Labor of Volunteer Moderators Online. Social Media + Society, 5(2). https://doi.org/10.1177/2056305119836778.

NARANJO DE LA CRUZ, Rafael. (2002). Los límites de los derechos fundamentales en las relaciones entre particulares: la buena fe. Centro de Estudios Políticos y Constitucionales, Madrid.

ORTEGA GIMÉNEZ, Cristina.(2024). El discurso del odio desde una perspectiva constitucional: cuando el castigo penal (casi) nunca sirve para proteger a personas vulnerables. *Cuadernos de RES PUBLICA en derecho y criminología*, (3),108–127. https://doi.org/10.46661/respublica.9544

PÉREZ LUÑO, Antonio Enrique. (2024). Desafíos a los que se enfrenta el discurso de los derechos. Derechos y Libertades, 50:95-108. https://doi.org/10.20318/dyl.2024.8234.

PEÑA JIMÉNEZ, Pedro José. (2021) Entre analogías y metáforas: el debate sobre la moderación de contenidos en las redes sociales. Revista de las Cortes Generales, 111:265-311. https://doi.org/10.33426/rcg/2021/111/1614.

SARAZÁ JIMENA, Rafael. (2008) Jueces, derechos fundamentales y relaciones entre particulares. Dissertation, Universidad de La Rioja.

SCHWEMER, Sebastian Felix. (2019) Trusted notifiers and the privatization of online enforcement. Computer Law & Security Review, vol. 35. https://doi.org/10.1016/j.clsr.2019.105339.

SUNSTEIN, Cass R. (2017) Republic: divided democracy in the age of social media. Princeton University Press, Nueva Jersey. https://doi.org/10.1515/9781400884711

SUZOR, Nicolas. (2017) Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. GigaNet. http://dx.doi.org/10.2139/ssrn.2909889.

TERUEL LOZANO, Germán Manuel. (2023) Libertad de expresión, censura y pluralismo en las redes sociales. In: Balaguer Callejón F (ed) and Cotino Hueso L (ed) Algoritmos y el nuevo paradigma regulatorio europeo. Fundación Manuel Giménez Abad, Zaragoza, p. 181-222.

VÁZQUEZ ALONSO, Víctor Javier. (2022) La censura «privada» de las grandes corporaciones digitales y el nuevo sistema de la libertad de expresión. Teoría & derecho, 32:108-129. https://doi.org/10.36151/td.2022.040.

WU, Tim (2003) Network Neutrality, Broadband Discrimination. Journal of Telecommunications and High Technology Law, 2:141-180. https://scholarship.law.columbia.edu/faculty_scholarship/1281/.