The Digital Services Act – An Appropriate Response to Online Hate Speech?

Pascal Schneiders & Lena Auler

Abstract

Online hate speech seems to permeate Facebook, X, Telegram, and the like, prompting increased national and supranational pushes for regulation of digital platforms. One of the most recent high-profile legislative frameworks is the Digital Services Act, which includes cross-sectoral and EU-wide moderation, transparency, and other due diligence obligations that are tiered according to the role, size, and impact of the online services. This chapter presents and critically analyses the measures raised in the Digital Services Act that are relevant to curbing hate speech. It concludes with recommendations for the future academic and regulatory approach to online hate speech.

1. Introduction

Social media platforms, such as Facebook, Instagram, and Reddit, have long since become central venues not only for maintaining relationships and seeking entertainment, but also for consuming and commenting on content (Newman, 2023). However, hopes that social media would evolve into arenas of deliberate discourse – if they ever existed beyond the small circle of a tech-savvy avant-garde – can rightly be described as dashed. Instead, there is now a widespread impression that a heated public sphere (Wagner, 2019), an outrage industry (Berry and Sobieraj, 2016), or even digital fascism (Fielitz and Marcks, 2019; Fuchs, 2022) prevails in the posts and comment sections of Facebook and Co. Hate speech, which is not a standardised legal term (Koreng, 2017; Valerius, 2020), but in social science usually refers to the "bias-motivated, hostile, and malicious speech aimed at a person or group of people because of some of their actual or perceived innate characteristics" (Cohen-Almagor, 2011, p. 1; see also Erjavec and Kovačič, 2012; Sponholz, 2023), seems to poison interactions on platforms and beyond (Bayer and Bárd, 2020; Udupa et al, 2021).

Indeed, content analyses have demonstrated hate speech's presence on social media despite the existence of content-moderation measures (Hestermann et al, 2021). Such speech represents only a minority of social media content (Siegel, 2020), and exists mostly in the form of stereotyping rather than the most drastic forms, such as incitements to violence (Paasch-Colberg et al, 2022). Nevertheless, many users are exposed to hate speech. An annual survey of internet users aged 14 and over in Germany shows that the proportion of respondents who had encountered hate speech online has remained consistently high for years (around 75%). Especially adolescents and young adults are exposed to hate speech (Landesanstalt für Medien NRW, 2023; see also Keipi et al, 2017). For those affected, especially younger people, hate speech has primarily psychological consequences, ranging from emotional stress and anxiety to depression (Keipi et al, 2017; Lee-Won et al, 2020). Furthermore, hate speech can silence vulnerable groups and demobilise them from participating in public life. By spreading hate speech and suggesting a or intimidating the majority opinion, highly active predominantly right-wing, networks can discourage people who are not themselves under attack from entering into the discourse and normalise negative stereotypes and radical views in wider circles (Das NETTZ et al, 2024; Gelber and McNamara, 2016). Not last, hate speech can incite others to make extremely uncivil statements or even to commit acts of violence (Müller and Schwarz, 2021; Williams et al, 2020).

It seems plausible that it is the specific platform logics that facilitate the emergence, dissemination, reception, and impact of hate speech (see also Recuero, 2024). That is, the affordances, rules, and algorithmic values of social media encourage low-threshold communication and networking, and incentivise exaggerated and emotional content – all of which serves to create a fertile environment for hate speech. In any case, "proprietors of spaces are responsible for the features of their spaces that present hazards by posing risks of harm if not managed" (Price, 2021, p. 260).

While hate speech has long been an issue that has received little political attention (Banks, 2010), politicians now never tire of insisting that the internet is "no lawless space" (see, for example, EPP Group, 2021; The Economist 2018; Cooper 2018). Against this backdrop, content moderation, understood as "the screening, evaluation, categorization, approval or removal/hiding of online content according to relevant communications and publishing policies" (Flew et al, 2019, p. 40; see also Art. 3 lit. t DSA), is becoming increasingly important. Soft law measures, which at the EU level

are essentially the "Code of conduct on countering illegal hate speech online" implemented in 2016 together with Facebook, Microsoft, Twitter (later, X), and YouTube (European Commission, 2016), and the Commission Recommendation (EU) 2018/334 of 2018 on measures to effectively tackle illegal content online, were clearly insufficient in the Commission's view for curbing hate speech. The European Commission (EC) saw a need for improvement in terms of transparency and feedback to users on the decisions about their notices. The Audiovisual Media Services Directive (AVMSD) (Directive 2010/13/EU), which prohibits the distribution of content that incites violence or promotes hatred, only applies on a sector-specific basis, e.g., to providers of video-sharing platforms, but not to text-based media or platforms. Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online, which entered into application in June 2022, is limited to illegal terrorist content.[1]

Accordingly, the EC presented the proposal for a Digital Services Act (DSA) (European Commission, 2020b) in December 2020 together with the proposal for a Digital Markets Act (DMA) (European Commission, 2020a).[2] The final version of the DSA (Regulation 2022/2065), which came into full force in February 2024, serves to update the 2000 Directive on electronic commerce (e-Commerce Directive) (Directive 2000/31) and significantly extends binding platform regulation. Some of the recitals, definitions, and procedures of the above-mentioned Code of Conduct and Commission Recommendations have been recognisably incorporated into the DSA (Cole et al, 2020). While the main features of the existing liability regime enshrined in the e-Commerce Directive remain "essentially the same" (Jaursch, 2021, N. 16; see also Cauffman and Goanta, 2021; Hofmann, 2023, p. 113), the DSA introduces detailed new transparency, moderation, and other due diligence rules, including risk assessments, au-

1 The regulation provides for hosting service providers to be obliged to apply measures to remove terrorist content from their services without delay. Authorities to be designated by the Member States (whether administrative, law enforcement, or judicial) are authorised to order hosting service providers to remove or disable access to terrorist content found to be illegal in a court or administrative decision within one hour throughout the EU (Art. 3 para. 1 Regulation [EU] 2021/784). In addition, service providers may be required to take further measures to prevent the public dissemination of terrorist content, such as the establishment of reporting mechanisms for users (Art. 5 para. 2b Regulation [EU] 2021/784).

2 For more information on the Digital Markets Act, see Chapter 6 'The Brave Little Tailor v. Digital Giants: A Fairy-Tale Analysis of the Social Character of the DMA' by Liza Herrmann.

dits, and research data access rules. In so doing, the DSA aims to ensure a harmonised, safe, predictable, and trustworthy online environment in which the fundamental rights enshrined in the EU Charter of Fundamental Rights (CFR) are "effectively protected" (Art. 1 para. 1 DSA) – this also includes the protection of the personal rights of those affected by hate speech (Kalbhenn and Hemmert-Halswick, 2021; Kapusta, 2024). Consequently, significant hopes are placed on the DSA to help curb hate speech. As early as October 2023, the EC sent X (formerly, Twitter) its first formal request for information under the DSA due to the spread of violent content and hate speech after the Hamas-led attack on Israel (European Commission, 2023a). In January 2025, a revised version of the Code of conduct on countering illegal hate speech online (the 'Code of conduct+') was integrated into the regulatory framework of the DSA. To date, the Code of conduct+ was signed and submitted for integration under the DSA by services such as Facebook, Instagram, LinkedIn, Snapchat, TikTok, Twitch, X and YouTube (European Commission, 2025b).The DSA may also have been motivated by the fact that, in recent years, some Member States have already made national progress in terms of new requirements for content moderation on digital platforms. For instance, Germany introduced the "Netzwerkdurchsetzungsgesetz" (NetzDG, Network Enforcement Act) (BGBl. I 2017, p. 3351), which came into force in January 2019, France implemented the "loi visant à lutter contre les contenus haineux sur internet" (loi Avia, "law aiming to fight against heinous content on the internet") (LOI n° 2020-766),[3] and Austria advanced the now-repealed "Kommunikationsplattformen-Gesetz" (Communication Platforms Act) (BGBl. I Nr. 151/2020), which came into force at the beginning of 2021. The NetzDG in particular, which may well have been the first law of its kind, has attracted global attention, been subject of controversial debate (Schulz, 2019, pp. 13–14), and has served as a source of inspiration for the DSA (Holznagel, 2021, p.123).

The DSA has been described as a "legislative mega-project" (Holznagel, 2021, p. 123) and a "constitution for the internet" (Geese, 2022). It is said to "represent the furthest reaching expansion of platform regulation in the OECD nations to date" (Cioffi et al, 2022, p. 828), potentially affecting the

3 In June 2020, the Conseil Constitutionel declared the law passed by the National Assembly in May of the same year to be unconstitutional, particularly because the one-hour deletion period imposed on the platforms for obviously illegal content constituted an unreasonable, unnecessary, and disproportionate interference with freedom of expression (Décision n° 2020-801 DC du 18 juin 2020, para. 8). The law was subsequently adapted to the court's requirements and published.

freedom of expression of millions of EU citizens and having a regulatory impact far beyond the Union's borders. Therefore, its measures against hate speech and suitability should be analysed all the more intensively (Latzer et al, 2019). In particular, experience with the NetzDG can help to understand the opportunities and risks of the content moderation measures against illegal content provided for in the DSA. The remainder of this chapter discusses the ways in which hate speech is dealt with within the DSA. First, the general regulatory approach of the DSA is discussed. Next, relevant provisions for platforms (in particular, the notice and action procedure), additional due diligence obligations for very large online platforms (VLOPs), and the transparency obligations contained in the DSA are presented. Subsequently, selected aspects of the regulation, including the privatisation of law enforcement and the effectiveness of content moderation in dealing with hate speech, are critically discussed. The chapter concludes with recommendations for the future regulatory treatment of hate speech.

2. Regulation of online hate speech in the DSA

2.1 Regulatory approach of the DSA for content moderation

In the context of platform regulation, particularly when it comes to the design of provisions for content moderation and dealing with hate speech, EU legislators face the challenge of harmonising the interests of the various stakeholders involved in digital communication. This proves to be a difficult balancing act, especially as there are multiple different interests in this context (Berberich, 2023, p. 130).

The fundamental rights of communication require the guarantee of open discourse as a basic prerequisite for a democratic society. On the one hand, the fundamental rights of users, who can invoke their freedom of expression and information when posting and consuming content (Art. 11 CFR), must be taken into account. At the same time, it must be ensured that users are adequately protected from the negative consequences of the dissemination of unlawful content, such as discrimination (Art. 21 CFR). Moreover, users must be guaranteed that, in case of an infringement, they can also take action in the digital communication space. On the other hand, due diligence obligations affect the services' freedom to conduct business (Art. 16 CFR). For their part, the services can also invoke the right to

freedom of expression. When drafting their terms of use, the providers can decide within the limits of their private autonomy which requirements they wish to set for the use of their services, and can thus also moderate unwanted (but not illegal) content on their platforms (Adelberg, 2022; Berberich, 2023, pp. 144–155).

The DSA attempts to address these complex interests by opting for a regulatory concept based on state–private co-regulation (Hofmann and Raue, 2023, p. 37). The legislator delegates the moderation of the content published on their platforms to the service providers, which seems to be without alternative considering the large amounts involved (Brauneck, 2024, p. 379). Indeed, questionable content can hardly be viewed and classified manually. The DSA also imposes a variety of due diligence obligations on platforms, which act as counterweights to the services' privileged liability and which the DSA can monitor and enforce by establishing a European supervisory structure. For example, providers of intermediary services are obliged to conduct content moderation "in a diligent, objective and proportionate manner" and to take the fundamental rights of the services' users into account (Art. 14 DSA). By limiting itself to procedural, content-independent requirements, the Regulation guarantees the protection of fundamental rights through procedures and a principle for procedural fairness (Berberich, 2023, p. 130).

2.2 Illegal content and hate speech

The DSA mainly targets *illegal* hate speech (Recitals 12, 62, 80, 87, 106 DSA), but does not define *what* content is illegal, just as the EC Directive does not define illegal activities. What is defined as illegal remains (for the time being) a matter for the Member States' or other EU legislation (Art. 3 lit. h DSA). However, there are (as yet) no legal definitions of hate speech as a legal concept in the EU Member States (European Commission, 2021). This means that hate speech – that is, the combination of a (supposed) group reference and the public, inflammatory defamation of this group or its (supposed) members (see Section 1) – does not necessarily have to be unlawful; it can also be contained in permissible expressions of opinion (Brugger, 2003). A binding framework for the definition and prosecution of serious forms of racist and xenophobic hate speech and crimes was established by the Council of the European Union in Framework Decision 2008/913/JHA of 28 November, 2008. In the report on the implementation

of the Council Framework Decision, the Commission clarifies that "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin", as well as "publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes" (Art. 1a, 1c Framework Decision 2008/913/JI) is to be treated as a (racist or xenophobic) criminal offence or *hate speech* in the Member States (European Commission, 2014). The conduct must be intentional and have a certain potential impact, i.e., be carried out "in a manner likely to incite violence or hatred against such a group or a member of such a group" (Art. 1c Framework Decision 2008/913/JI). At the end of 2021, the EC asked the Council to introduce an EU-wide definition of hate speech and include it in the list of so-called EU crimes. The latter are crimes of a particularly serious nature with a cross-border dimension, as set out in Art. 83 para. 1 of the Treaty on the Functioning of the European Union (TFEU) (European Commission, 2021). This would be accompanied by EU-wide minimum rules on the definition of criminal offences and penalties. However, the process has been stalled in the Council since 2022 (European Parliament, 2024).

2.3 Provisions for hosting services, online platforms, and VLOPs

The DSA does not impose completely new measures on digital platforms. They have been active in self-regulation for many years, and automatically and proactively moderate large-scale, third-party content (Gorwa et al, 2020; Klonick, 2018.). Thus, the DSA formalises practices and standards for curbing hate speech. The horizontal obligations for all online intermediaries listed in the DSA are graded according to the scope of the digital services, which are divided into: 1) intermediary services, 2) hosting services, 3) online platforms, and 4) VLOPs and very large search engines (VLOSEs) ("pyramid-model"; Hofmann and Raue, 2023, p. 33). While intermediary services merely pass on information provided by users or store it temporarily for the sole purpose of (efficient) transmission, hosting services store the information provided on behalf of their users (Art. 3(g) DSA). Hosting service providers that store information and make it available to the public on behalf of a user (e.g., app stores and social media platforms) are considered online platforms (Art. 3(i) DSA). VLOPs have a significant reach in the EU (by definition, at least 45 million monthly active users or, in

case of a decreasing or increasing population, 10% of the EU population) (Art. 33 para. 1 and 2 DSA) and have a particular social and economic impact (Justification and Recital 79 DSA). Accordingly, they must fulfil the most comprehensive catalogue of obligations, including internal risk assessments, external audits, and data exchanges with authorities and researchers. To date, the Commission has designated two VLOSEs (Bing and Google Search) and 25 VLOPs, including Instagram, YouTube, TikTok, Facebook, X, and LinkedIn (European Commission, 2025b).

The content moderation measures formulated in the DSA include mechanisms for notifying illegal content as well as, in a broader sense, complaint procedures, out-of-court dispute settlement bodies and, last but not least, service providers' obligation to report suspected serious offences to the competent authorities. These decisions by the platforms should be swift, transparent, and contestable for all parties involved (de Streel et al, 2020, p. 79). In the following, the measures contained in the DSA – from liability obligations for user-generated content to due diligence obligations concerning the design and operations of services – are considered in greater detail. First, it is important to discuss when platform providers are responsible for the user-generated content disseminated on the platforms.

2.3.1 Notice and action procedure

At the heart of the DSA's content moderation measures is Art. 16, which requires hosting services and (VL)OPs to establish procedures for individuals or institutions to provide notices for content they consider to be illegal. The notice and action mechanism provided for in Art. 16, especially the presumption of knowledge in para. 3, is closely linked to the liability regime in Chapter II of the Regulation (Gerdemann and Spindler, 2023, p. 8; Raue 2023, p. 290). The liability privileges established in Chapter II exempt providers from responsibility for third-party content (Hofmann, 2023, pp. 129–131). Art. 8 DSA clarifies – in line with the liability concept of the e-Commerce Directive – that the providers of intermediary services are not subject to any proactive precautionary and investigation obligations regarding illegal content. This privilege is based on the notion that, due to the large amount of content that is distributed on platforms, providers are unable to check every single piece of content individually. Without the privilege, business models of online platforms could be jeopardised (Hofmann, 2023, p. 184). They can only be obliged to block or remove content as soon as they become aware of illegal activity or content. Knowl-

edge may be obtained, for example, through a notification by users or other organisations in the sense of Art. 16 (para. 3). The notice and action mechanism thus compensates for a weakness in liability by implementing a procedural reporting obligation and counterbalancing the exemptions from liability (Legner, 2024, p. 106; Raue, 2023, p. 289). Service providers are obliged to process all notices and decide on them "in a timely manner" (Recital 52 DSA). If they use automated means for processing or decision-making, the person or organisation that has submitted the notice must be informed of this (Art. 16 para. 6). In contrast to Regulation (EU) 2021/784 or the NetzDG, the DSA does not set any time limits for the processing period. The NetzDG required social network providers to remove or block access to "manifestly unlawful" content reported by users or complaints bodies within 24 hours. However, the signatories of the Code of conduct+ committed to review the majority (at least 50%) of hate speech notices from so-called (trusted flagger-like) Monitoring Reporters within 24 hours (European Commission, 2025a).

According to the DSA, service providers must inform users and give reasons when users are affected by the following restrictive moderation decisions that are imposed on the ground that the user-generated or -distributed information is illegal or incompatible with the terms and conditions: a) any restrictions of the visibility of specific information items, such as the removal, demoting, or blocking of content; b) restriction of monetary payments; c) suspension or termination of the provision of the service in whole or in part; and d) suspension or termination of the user's account. The statement of reasons shall be provided at the time of the removal or blocking at the latest (Art. 17 para. 2 DSA). If individuals or entities abuse the notice and action mechanisms by frequently submitting obviously unfounded notices – i.e., for the purpose of silencing marginalised groups (Duffy and Meisner, 2023) – the platform providers shall suspend the processing of the notices and complaints (Art. 23 para. 2 DSA).

Likewise related to content moderation, Art. 7 of the DSA introduces the so-called "good Samaritan privilege". It means that providers benefit from the liability privileges of the DSA if they conduct voluntary investigations on their own initiative or take other measures to detect, identify, remove, or disable access to illegal content. The provision clarifies that voluntary investigations do not automatically establish an active role that would remove the liability privileges. It is intended to prevent platforms that want to proactively prevent infringements with good intentions from being penalised (Koehler, 2024, p. 118; Kuczerawy, 2021). Providers should

not be deterred from taking voluntary measures. However, they must ensure that an objective, non-discriminatory, and proportionate procedure is in place that takes into account the rights and interests of all parties involved (Hofmann, 2023, p. 128). In this context, according to Recital 26, providers should take protective measures against the unjustified removal of lawful content. To that aim, providers should, for example, take reasonable measures to ensure that, where automated tools are used to conduct such activities, the relevant technology is sufficiently reliable to limit to the maximum extent possible the rate of errors.

In addition, the DSA encourages cooperation between platforms and third parties – so called trusted flaggers – in detecting and notifying – and only of – illegal or unlawful content. Trusted flaggers receive notices of illegal content from users, but can also search online platforms for illegal content themselves. They are awarded by the Digital Services Coordinators (DSCs)[4] upon request and have to be independent from online platforms, but not necessarily from state authorities (Art. 22 para. 2 DSA). Accordingly, trusted flaggers can include industry organisations, authorities as Europol, or the criminal content units of national law enforcement agencies, including the National Internet Referral Unit at the Federal Criminal Police Office (Bundeskriminalamt, BKA) in Germany[5] (Recital 61 DSA).[6] Trusted flaggers are required, among other things, to have expertise in dealing with illegal content, to represent collective interests, and to carry out their activities diligently, accurately, and objectively. Notices submitted by trusted flaggers should be given priority in the platforms' content mod-

4 The DSCs are appointed by Member States and are responsible for various issues relating to the application and enforcement of the Act (Art. 49(2) DSA). Together with the Commission, coordinators and – depending on specific Member State provisions – additional competent national authorities form the DSA's oversight structure. The DSCs should fulfil their tasks impartially and independently, i.e., they must not take instructions from other authorities (Art. 50 para. 2 DSA). In extreme cases, they are authorised to take interim measures to prevent the risk of serious harm (Art. 51 para. 2e DSA). In addition, a European Board for Digital Services, which advises the coordinators and the Commission, will help ensure the uniform application of the act (Art. 61 para. 2a DSA).

5 Internet Referral Units (IRUs) actively search the internet for criminal or extremist content. On the potential for abuse of the EU-wide IRUs, see Chang (2018), who expressed the concern that "IRUs are setting a dangerous precedent of state-initiated, privately-enforced, and extra-legal censorship that could be abused to limit speech that is neither genuine incitement to violence nor terrorism" (p. 124).

6 As part of the efforts to combat terrorist content, such service providers as Google and YouTube have already awarded IRUs' Trusted Flagger status (Chang, 2018).

eration decision (Art. 22 para. 1 DSA). That is, the decision on the content of illegal content remains with the platforms (Ruschemeier, 2024). If an investigation by the DSC reveals that a trusted flagger no longer fulfils its requirements, the DSC can revoke that status. This is the case if, for example, the trusted flagger demonstrates a lack of expertise, diligence, and objectivity, or frequently submits inaccurate or unsubstantiated notices. Investigations can be made ex officio – that is, a regulatory authority may initiate an investigation on its own without a complaint having been filed – or in response to information from third parties regarding the behaviour of trusted flaggers (Art. 22 para. 6–7 DSA). As such, the DSA does not provide for the permanent and regular watching of the watchmen; instead, the monitoring of trusted flaggers is largely based on the observation of their work by third parties.

It should be noted that digital platforms have been working with trusted flaggers for years, among other reasons, because of the Code of Conduct on countering illegal hate speech online (see Section 1). In Germany, trained organisations participating in YouTube's Priority Flagger programme include, for example, the BKA, several state criminal investigation offices, jugendschutz.net, the German Association for Voluntary Self-Regulation of Digital Media Service Providers (Freiwillige Selbstkontrolle Multimedia-Diensteanbieter, FSM), media state authorities, and non-profit associations. Their notices of content that violates YouTube's community guidelines are given priority (Google and Youtube, 2019).

2.3.2 Complaint and redress mechanisms

According to Art. 20 para. 1 of the DSA, users can dispute: 1) the removal, blocking, or demoting of content deemed illegal or incompatible with the general terms and conditions; 2) the suspension or termination of the service; 3) the suspension or termination of the user account; and 4) the suspension or restriction of monetisation options by the service provider. This should be possible via an internal complaint-handling system for a period of at least six months after the moderation decision. It should be possible to lodge a complaint regardless of whether the moderation decision was made proactively by the platform or in response to a notice from a user or trusted flagger. Online platforms must process complaints promptly. Furthermore, users should be able to appeal to an independent out-of-court dispute settlement body certified by the Member State's DSC (Art. 21 DSA). If the dispute settlement body decides in favour of the user, the online

platform will bear all fees and costs. If the decision is unfavourable to the user, the user shall bear only his own fees and costs (Art. 21 para. 5). Users are still free to seek legal protection in court against the online platform's decision to restrict an information piece, payments, account, or its service.

2.3.3 Cooperation with authorities

Moreover, the DSA requires service providers to cooperate with the authorities. This includes reporting obligations and complying with official orders. For example, the DSA obliges providers of hosting services to contact the respective Member State's law enforcement or judicial authorities if they have reasonable grounds to suspect that a serious criminal offence "has taken place, is taking place or is likely to take place". In this context, the provider shall make all relevant information available to the authorities (Art. 18 para. 1), including, *where relevant*, information required to locate and identify the respective user of the service (Recital 56).

Further to reporting, the DSA gives national judicial or administrative authorities the option of issuing reasoned orders to providers of intermediary services, including foreign providers, to provide information about individual users (Art. 10) or to take action against certain content found to be illegal (including cross-border content) (Art. 9). The EU Regulation on addressing the dissemination of terrorist content online (Regulation 2021/784)[7], which entered into force in June 2022, contains a similar mechanism for taking action against certain types of illegal content. The orders addressed to providers might also be aimed at preventing the reappearance of illegal content, but without imposing a general monitoring obligation (Recital 30). When determining the territorial scope of the order, the authorities are required to weigh up the interests at stake and, in particular, to consider the rights enshrined in the EU Charter of Fundamental Rights, such as the freedoms of expression and information (Recital 36 DSA). The official information and moderation orders must be documented in the transparency reports (Art. 15 para. 1a DSA). In case providers do not comply with the orders, the DSA itself does not lay down any consequences, with enforcement instead being a matter of national law. This stands in contrast with the information obligations under Art. 9 para. 1 and Art. 10

7 For more information on this Regulation, see Chapter 7, 'Eyes shut, fingers crossed: the EU's governance of terrorist content online under Regulation 2021/784' by Valerie Albus.

para. 1 DSA, which can be enforced by means of the regulation, such as through fines (Recital 32 DSA; see also Hofmann, 2023, p. 196, 201). Providers merely have to state that they have received the order and how they have complied with it (Art. 9 para 1, Art. 10 para. 1).

2.4 Additional due diligence obligations for providers of VLOPs and VLOSEs

VLOPs and VLOSEs must fulfil special due diligence obligations, including risk assessments, risk mitigation, audits and data access. The supervision and enforcement of these obligations is the sole responsibility of the EC (Art. 56 and Art. 2 DSA).

2.4.1 Risk assessment and mitigation

Risk assessments relate to systemic risks arising from the design, functioning, use or misuse of the services (in accordance with Art. 34 para. 1 DSA). Systemic risks include: a) the dissemination of illegal content (Recital 80 gives the example of "illegal hate speech"); b) adverse effects of the service on fundamental rights (including the fundamental rights to human dignity and to freedom of expression and information); c) negative effects on democratic and electoral processes, social debate, and public safety; and d) negative effects in relation to gender-based violence, the protection of public health and minors, and serious negative consequences to a person's physical and mental well-being. Risk analyses must be conducted by the platforms themselves ("first party audit"; Meßmer and Degeling, 2023), proactively and on an annual basis, and before the introduction of new, critical functionalities (Art. 34 para. 1 DSA). Determining the extent to which hate speech constitutes a systemic risk arising from the design, operation, or use of VLOP/VLOSE services is thus initially the responsibility of the platforms. In doing so, they must pay particular attention to: 1) the terms and conditions, 2) the content moderation systems, and 3) the design of the algorithmic recommender systems. The procedure and criteria of the analysis are not predetermined in more detail. With the "risk management framework" (European Commission, 2023a), the EC proposed a methodology for risk assessment and mitigation (however, in the context of Russian disinformation campaigns and not in relation to hate speech). Accordingly, a distinction can be made between qualitative and

quantitative risk indicators. The quality of a risk posed by a particular type of content is assumed to be a function of a speech's context, the speaker's position or intent, the content or form of the speech, the reach, size, and characteristics of the audience, and the likelihood of harm. Quantitative measures comprise the audience's size and exposure to and engagement with the content, the content's prevalence, and the influence of algorithmic promotion (European Commission, 2023a).

If the VLOPs identify internal systemic risks, they must take reasonable, proportionate, and effective measures to mitigate them. The DSA lists a number of non-exhaustive measures in this regard, including the adaption of the terms and conditions, internal decision-making processes, the design, features, or functioning of their services, the advertising systems, the algorithmic recommender systems, and the content moderation processes (Art. 35 para. 1 DSA). In particular, platforms could adapt the responsiveness to user notices, the speed and consistency of removal and labelling, and the de-amplification of illegal or otherwise harmful content (algorithmic down-ranking, the removal of recommendation, searchability, and/or monetisation) (European Commission, 2023a). Moreover, VLOPs can cooperate with trusted flaggers to reduce systemic risks (Art. 35 para. 1g). The DSA does not concretely specify how platforms should proceed, thereby enabling VLOPs to try out different risk mitigation practices. However, the EC may furthermore require the application of preventive and remedial crisis response measures to assess threats and related measures (Art. 36), and request (VL)OPs to participate in the development of codes of conduct for risk reduction (Art. 45 DSA). Correspondingly, the EC announced that adherence to the Code of conduct+ may be considered as an appropriate risk mitigation measure for VLOPs and VLOSEs (European Commission, 2025c).

The internal risk analyses and mitigation plans must be assessed for compliance by independent organisations ("second-party audit"; Meßmer and Degeling, 2023) at least once a year (Art. 37 DSA). The audit organisations are commissioned by the service provider. The EC has issued a Delegated Act (2024/436) with rules on independent audits to assess VLOPs' and VLOSEs' compliance with the DSA. According to Art. 290 TFEU, the Commission can use delegated acts to supplement or amend existing legislative acts. The aforementioned Delegated Act provides auditors with fairly comprehensive access to information on procedures and processes, decision-making structures, IT systems, data sources, algorithmic systems, information technology systems, testing environments, personnel, and in-

ternal compliance procedures (Art. 5 Delegated Act). It specifies audit procedures, defines minimum standards, and seeks to allow a certain degree of comparability of the reports. However, it does not specify any methods or quality criteria according to which the audits are to be conducted. The services were to be audited for the first time at the end of August 2024.

2.4.2 Data access

Access to platform data by independent research institutions is essential for assessing the extent and impact of hate speech (King and Persily, 2019; Rieder and Hofmann, 2020; Stark et al, 2020). Art. 40 para. 4 of the DSA – which was extensively amended during the legislative process – provides for private non-public access for "vetted researchers". This makes the DSA the first EU law to enable mandatory data access (Jaursch and Lorenz-Spreen, 2024). The possibility of data access is linked to the condition that the research contributes to: 1) the detection, identification, and understanding of systemic risks (Art. 34 para. 1) and to 2) the assessment of the adequacy, efficiency, and impact of risk mitigation measures (Art. 35). The draft Delegated Act (DDA) laying down the technical conditions and procedures under which VLOPs and VLOSEs are to share data mentions a variety of data that allow to study systemic risks. Among these are user-related data such as profile information, relationship networks, individual-level content exposure and engagement histories; interaction data such as comments or other engagements; data related to content (personalised) recommendations, and data related to content moderation and governance (Recital 12 DDA). This should also allow for studies on the role of platform logic and algorithmic recommendations in the dissemination of hate speech. Platforms shall make available an overview of the data inventory of their services easily accessible online, including examples of available datasets and suggested modalities to access them (Art. 6.4 DDA). Such modalities may be, among others, data transfer to the vetted researchers, and a transmission of the data to and storage in a secure processing environment which are to be operated by data providers themselves or by a third party (Recital 16 DDA). How data access is organised in detail is, to some extent, up to the platforms. This includes how the application programming interface (API) should be designed or in which format data should be made accessible (Van Drunen and Noroozian, 2024). As a first step, (groups of) researchers have to submit an application for vetted researcher status to the DSC where the platform(s) of interest is/are based or to the DSC of

the research organisation's Member State. In the course of this process, the researchers submitting the application must address the specific research for which they consider data access to be necessary (Art. 40 para. 8). One of the admission requirements is that the researchers must be affiliated to scientific (not exclusively academic) research organisations and do not pursue commercial interests. This may also include civil society organisations. Researchers can also be non-EU based (Albert, 2024). Within 21 days from the receipt of a data access application that fulfills all prerequisites (such as information about funding, and a description of the research project and planned methodology; Art. 8 DDA), the DSC where the main establishment of a provider is located will decide whether to transmit a reasoned request to the relevant VLOP or VLOSE and inform the researcher of its decision (Art. 7 DDA). The DSC also determines the modalities according to which access to the data is to be granted by the platform. A key factor here is how sensitive the data is (Art. 9 DDA). The platform then has 15 days to ask the DSC for amendments to the request. This is only possible if the service provider considers that it cannot comply with the request due to a lack of access to the data or due to concerns about the security of the service or the protection of confidential information (Art. 40 para. 5). However, the platform provider must offer alternatives on how access to the requested (or other) data can be granted (Art. 40 para. 6). This makes it more difficult to evade data access by invoking business secrets. The DSC decides on the request for amendment within a further 15 days. DCSs may consult independent experts before formulating a reasoned request or taking a decision on an amendment request (Art. 14 DDA).

In addition to the data access, platforms should provide vetted researchers with the relevant metadata and data documentation (such as codebooks) so that they can cope with the data (Recital 26 DDA). In future, a data access portal hosted by the EC will provide a public overview of all reasoned requests sent by the DCSs (including not successful ones). VLOPs and VLOSEs are also obliged to provide immediate access (e.g., without having to contact a supervisory authority) to (real-time) data that are *publicly* accessible in their online interface by researchers and used to investigate systemic risks (Art. 40 para. 12). This can be interpreted as a "right to scrape" (Klinger and Ohme, 2023). Said publicly available data may, for example, include data "on aggregated interactions with content from public pages, public groups, or public figures, including impression and engagement data such as the number of reactions, shares, comments from recipients of the service" (Recital 97). As it is not limited to researchers who

are affiliated to a research organization, NGOs and journalists could also make use of this right.

2.5 Transparency obligations

Art. 15 DSA stipulates that providers of intermediary services and of (VL)OPs must disclose the measures they have taken regarding notices submitted in accordance with Art. 16 or on their own initiative in a transparency report published at least once a year. Shorter reporting cycles of six months apply to VLOPs (Art. 42 para. 1). In the transparency reports, the service providers must indicate whether the moderation decisions were made on a legal basis or according to their own general terms and conditions (Art. 15 para. 1b). Online platforms must also state, among other things, the extent to which automated means are used for content moderation, as well as their precision (Art. 15 para. 1c, 1e). As the DSA thus formulates transparency obligations regarding both illegal content and content that does not comply with the general terms and conditions, platforms cannot escape regulation by (increasingly) moderating according to their own standards. Such an effect was observed after the introduction of the NetzDG in Germany (Kalbhenn and Hemmert-Halswick, 2021).

Further to the reporting obligations relating to the user notices procedures, VLOPs and VLOSEs must file publicly available reports outlining the results of the first-party audits on risk analysis, but only after they have been audited by independent organisations (Art. 42 para. 4a). They also have to report on the risk mitigation measures they have been recommended by audit organisations and on those they have implemented (Art. 42 para. 4b, 4d). The second-party audit reports (on the service providers' compliance with the DSA regulations on risk assessment and mitigation) must be published within three months of receipt from the auditing organisation (Art. 42 para. 4c). Last but not least, vetted researchers who have been granted access to data are obliged to make their research results available free of charge "within a reasonable period after the completion of the research" (Art. 40 para. 8g).

In light of the above, it remains to be seen how effective and appropriate the DSA's measures are for combatting hate speech. Certainly, this question can only be answered after a longer period of time, when provisions have been fully implemented and empirical legal studies have been conducted.

However, some evaluation criteria and critical aspects can already be discussed.

3. Evaluation of the Regulatory Measures

3.1 Legitimacy and accuracy of content moderation

In terms of evaluation criteria, regulatory measures should first and foremost be legitimate (i.e., in line with fundamental rights). Furthermore, they should be suitable for solving the identified problem – that is, the low-threshold, significant generation and dissemination of hate speech, in particular on social media platforms. This solution should be appropriate, i.e. it should consider and balance different fundamental rights.

In the run-up to the implementation of the NetzDG in Germany, there was fierce criticism of an alleged privatisation of law enforcement (Pohlmann et al, 2023). This accusation can also be applied to the DSA (Cauffman and Goanta, 2021). In the first instance, it is the service providers who interpret the law and decide which reported content is litigable and should thus be removed or blocked (however, without the service providers taking over the prosecution and the final decision still being made by the courts; Hong, 2022). Other concerns that have been raised in connection with the NetzDG relate to the restriction of freedom of expression through over-removal by service providers (Mchangama and Fiss, 2019). After all, the NetzDG would create economic and regulatory incentives for service providers to remove content in cases of doubt in order to avoid reputational damage or fines (Buiten et al, 2020). With regard to the DSA, it's "good Samaritan" clause increases the risk of overblocking, as it encourages platform providers to engage in proactive content moderation – with the latter being challenging for external observers to comprehend (Kuczerawy, 2021).

Furthermore, the formalisation of digital platforms' content moderation obligations is linked to an increase in their opinion power (Helberger, 2020; Senftleben, 2024). However, the relatively low proportion of removed or blocked content in all NetzDG complaints supports the assumption that, to date, the extent of overblocking has tended to be overestimated (Kohl, 2022). Solid proof of over- or underblocking would require an in-depth and systematic legal review of notified content, including supposedly obviously illegal content, which is not practically feasible. Due to its systemic regula-

tory approach, there are no *direct* sanctions against under- or overblocking in the DSA itself. However, that said, the complaint and review procedures set out in the DSA could reduce the risks of the latter (Buiten et al, 2020; Cornils, 2020).

Apart from this, the independent audits to be conducted by VLOPs to ensure compliance with their obligations (Art. 37 DSA) should reduce the risk of over- or underblocking. In this context, it should also be examined whether the platform providers have carefully and objectively processed the notices received via the internal complaint-handling mechanisms. Furthermore, VLOPs must also explicitly consider the possible (negative) effects of content moderation on, among other things, freedom of expression and information when assessing systemic risks (Art. 34). However, it is up to the platforms themselves to define, assess, and address systemic risks (Griffin, 2023). Another critical point is that there are no minimum standards for conducting audits.

3.2 Involvement of state authorities

As far as the information provision obligations of platforms towards law enforcement or judicial authorities are concerned, consistent prosecution of illegal offences online is generally desirable. Otherwise, perpetrators are unlikely to change their minds (Kettemann, 2019). Failure to do so may give the impression of a lack of interest in enforcing norms, which in turn may lower the inhibition threshold for further hate speech (Rüdiger, 2019). Enforcement of legislation could and should also be enhanced by the establishment of additional prosecutors specialising in hate speech and related phenomena.

The authorisation of national authorities to order (EU-wide) action against allegedly illegal content, as provided for in the DSA, has the potential for instrumentalisation or abuse by state actors. For example, there could be boundary shifts regarding politically unpopular content and what can or cannot ultimately be removed by order. Limiting this risk of abuse, it should be noted that what is defined as illegal must accord with EU law (Art. 9 para. 1). More concerning is the risk that authoritarian governments outside the EU will use the measure in question to legitimise their own laws in the supposed fight against terrorism, extremism, fake news, or hate speech (Chang, 2018). For instance, Turkey, Russia, Belarus, and Malaysia have introduced legislation similar to the NetzDG, but with the aim of cen-

soring content critical of their regimes rather than protecting freedom of expression (Mchangama and Fiss, 2019; Reporters without Borders, 2017).

3.3 Effectiveness of content moderation

Regarding the suitability or effectiveness of co-regulatory content moderation measures in curbing hate speech, only little evidence has thus far been produced (Courchesne et al, 2021). Most studies relate to the NetzDG in Germany or "deplatforming", that is, the removal of one's account on social media for breaking platform rules. For example, Hestermann et al (2021) found a decline in hate comments between the study periods of January 2018 and July to November 2020, although this observation cannot be clearly attributed to the introduction of the NetzDG. Andres and Slivko (2021) conducted a quasi-experimental study on the effect of content moderation. They analysed Twitter posts published by followers of the populist and far-right parties AfD in Germany and FPÖ in Austria on the topics of religion and migration between July 2016 and June 2019. Their analysis of the automatically determined, multidimensional hate speech intensity of the tweets showed that the amount and intensity of hate speech decreased moderately among AfD followers after the NetzDG came into force in January 2018, but not among FPÖ supporters. This speaks in favour of the effectiveness of the NetzDG. Moreover, case studies on deplatforming show that blocking access to accounts, and thus to their content, can prevent the dissemination of hate speech (Ali et al, 2021; Bodden et al, 2023; Fielitz and Schwarz, 2020; Hammer et al, 2021). In this way, hate groups are deprived of the infrastructure to recruit and mobilise members, organise internally, disseminate their content, finance their activities, and harass minorities or dissenters (Rogers, 2020).

3.4 Data access

Access to data is fundamentally relevant to policy and research, not least to provide regulators with a broader, previously fragmented evidence base on the spread, impact, and containment of hate speech and other phenomena on social media, and the role of platform logics in this context. Previously, all major platforms have attempted to restrict or prevent data donation and

scraping. For example, in summer 2023, X sued the non-profit organisation Center for Countering Digital Hate, which had conducted research on the dissemination of hateful content on social media, accusing them of having "unlawfully" scraped data from X (X Corp. v. Center for Countering Digital Hate Inc., 2023). In autumn 2023, X ended free access to its API for researchers (Kupferschmidt, 2023). In August 2024, the CrowdTangle analytics tool was discontinued (Meta, 2024a). It allowed trending content to be detected, as well as how often a link was shared and who shared it. Meta's Content Library (Meta, 2024b) neither provides access to news media nor offers the same research functionalities as CrowdTangle (Coalition for Independent Technology Research, 2024). Other collaborations initiated by the platforms, such as Facebook's Social Science One Project or its Ad Library, have also been heavily criticised by researchers due to very limited access and incomplete data. As Meta's depreciation of CrowdTangle deprives researchers and journalists of real-time election monitoring tools, which could impair the ability to track misinformation and disinformation, the EC launched formal infringement proceedings against Meta in April 2024 (European Commission, 2024b).[8] The EC has also already opened formal investigative proceedings against X and TikTok due to shortcomings in giving researchers access to publicly accessible data. It is to be welcomed that the EC seems willing to enforce the DSA's data access regimes.[9]

However, there are still some open questions and points of criticism regarding the procedures and access modalities (for a more detailed analysis, see Seiling et al, 2024). For example, whether the application for vetted researcher status, once successful, must be resubmitted for each research project is still unclear. Likewise, the procedure for cross-platform analyses –

8 After initiating formal proceedings, the Commission carries out an in-depth investigation and gathers evidence, for example by sending additional requests for information, conducting monitoring actions, interviews, inspections and requesting access to algorithms. In addition, the Commission may take further enforcement steps. The DSA does not set a legal deadline for concluding formal proceedings, which depends on various factors, including the case's complexity and the company's cooperation.

9 In this context, however, the fact that the EU Commission is in charge of supervising VLOPs and VLOSEs poses a risk – in addition to the lack of state neutrality: As the executive body of the EU, it can be put under political pressure, meaning that platform regulation can become a geopolitical bargaining chip. This is already becoming apparent in the trade dispute between the EU and the US. During the election campaign, US Vice President J. D. Vance threatened the EU with making further support for Ukraine in the Russian war of aggression dependent on whether the Commission would discontinue the ongoing proceedings against X (Scheer et al,. 2025).

whether several separate applications for data access are to be submitted to the relevant DSC – has yet to be specified. Last but not least, the delegated act does not provide researchers with any clear remedy if the data received does not conform to quality standards.

4. Conclusion

4.1 Starting point and the DSA's approach

Online hate speech is a problem that affects millions of EU citizens and has negative consequences not only for individuals online and offline, but also for society as a whole. It does not only constitute an insult or group-related devaluation of people, but also suppresses their freedom of expression and can incite others to violence. Online hate speech on social media has reached problematic levels of visibility *despite* the moderation efforts of platform providers according to community standards (which sometimes go beyond legal definitions of criminal offences; Liesching, 2021, pp. 106–107), social media editorial teams, and existing national regulations (e.g., Germany's NetzDG). This suggests that platforms are not consistently tackling hate speech, implying that (further) external or co-regulatory measures are needed (Buiten et al, 2020) – or that a significant proportion of hate speech is not considered unlawful or perceived as violating the platforms' community standards. The extent of (at least not illegal) hate speech is likely to increase in the future now that Facebook has ended its cooperation with fact checkers and reduced community standards (Stippler et al., 2025). This has already been demonstrated on X (Hickey et al., 2023; Arun et al., 2024).

Against this background, the DSA is an important legislative project in the EU to strengthen incentives to curb hate speech and impose standardised and binding complaint, deletion, objection, reporting, and data access obligations on digital platforms. As described earlier, the EC has already launched several formal investigative proceedings against VLOPs. For example, in December 2023, it announced formal infringement proceedings against X on the basis of suspected breaches in its data access obligations, failure to counter the dissemination of illegal content, and deceptive design practices (European Commission, 2023b). In January 2024, the EC sent formal requests to 17 VLOPs and VLOSEs to provide more information on the measures they have taken to comply with the obligation

to provide researcher access to publicly available data (European Commission, 2024c). In February, two days after full DSA implementation, the EC announced similar proceedings against TikTok for potential breaches in protecting minors against the platform's potentially "addictive design", advertising transparency, and data access for researchers (European Commission, 2024d). However, the DSA is no *constitution for the internet* (see above), as its amendments are too incremental. The DSA establishes no specific standards for dealing with hate speech (or disinformation or illegal content). Rather, the aim is to formalise and standardise previously self-regulatory content moderation processes and the associated legitimate interests and considerations. It helps to increase the accountability of platforms and empower a critical public through regular transparency reports, risk and countermeasure assessments, independent audits and reports on moderation decisions, obligations to provide reasons, and opportunities for objection (Buchheim, 2022).

In order to defend the freedom of expression that hate speech threatens, the DSA relies on cooperation between users, notice centres, and other trusted flaggers, authorities, and platforms. These actors are involved in different phases of hate-speech management, from identifying and moderating hate speech to sanctioning hate speech disseminators and structurally adapting platforms. This distributes the responsibility for a discourse arena free from hate speech across several shoulders instead of shifting it unilaterally (e.g., in the form of general monitoring) to individual actors while simultaneously exempting others (Bryson, 2023; Buiten et al, 2020; Griffin, 2023). The DSA is concerned with procedural improvements (of reporting and objection options, transparency, and compliance) and leaves the definition of illegal content to Member States themselves. To this end, the DSA performs a balancing act between effective (rigorous deletion by platforms) and legitimate (involvement of users) content moderation. For example, the DSA includes complaint mechanisms and a certified out-of-court dispute settlement body that allow users to challenge the content moderation decisions of digital platforms. In this context, it is irrelevant whether the moderation decision was taken proactively by the platform or was triggered by a user's notice. It would be useful to extend these mechanisms to cover not only the decision by a platform to *remove*, but also the decision to *retain* notified content. Moreover, the DSA also contains comprehensive reporting obligations. For example, intermediary services and online platforms must document whether they have made moderation decisions on a legal basis or

as per their own terms and conditions. Moreover, VLOPs have to disclose the results of first- and second-party audits.

4.2 Risks and opportunities in relation to the DSA

Of course, the implementation of the DSA is not without risk: the Act formalises the fact that it is the hosting service providers and online platforms that decide in the *first instance* what content is illegal. While independent state courts will, naturally, continue to make final decisions on the legality of content, taking legal action is unlikely to be attractive for the majority of users, meaning that they will often accept the provider's decision, and thus the initial decision will effectively be the final decision (Raue, 2023, p. 345). This leads to the criticism that due diligence obligations would turn platforms into "quasi-judges" (Spindler, 2017, p. 481; Berberich, 2023, p. 173). This does not mean that private companies are not allowed to do so. Rather, it is the lack of transparency in content moderation that is problematic (Heldt, 2019). The inability of independent third parties to scrutinise the moderation decisions of platforms means that the possibility of overblocking cannot be excluded. Already marginalised groups (e.g., sex workers and abortion rights activists) are particularly vulnerable to overblocking (Appelman, 2023; Haimson et al, 2021). At this point, insight into the specific community standards and access to data for independent research institutions is crucial for identifying and evaluating any systematic over- or underblocking. To date, the content moderation measures of social media platforms according to their own community standards have not been transparent. However, it is important to determine where, and on what basis, the red line for hate speech is drawn, as the accuracy and precision of content moderation measures have implications for effective freedom of expression. Encouragingly, the research data access regime prescribed by the DSA (Art. 40) should allow for analyses on the precision of (automated) content moderation measures taken by, and other systemic risks associated with, platforms.

4.3 Implications

In this context, data access should be free of charge and the data should be easily accessible and findable, machine-readable, able to be structured

(e.g., by outlet), interoperable, and replicable. Metrics should be easy to understand (Democracy Reporting International, 2024; Ranaivoson and Domazetovikj, 2023; Specht-Riemenschneider, 2021). Data preparation should meet uniform, comparable standards, and data pools should be accessible in their entirety (Klinger and Ohme, 2023).

Moreover, data access regimes should be adapted to the (dynamic) needs of researchers (Van Drunen and Noroozian, 2024). On this basis, there is a need for cross-platform, continuous studies focusing on: 1) the reach and frequency of exposure to different degrees of hate speech, as well as their origins and evolution; 2) the differentiated effects of different degrees of hate speech at the individual and societal levels; 3) algorithmically induced radicalisation effects; and 4) the principles and accuracy of content moderation measures to curb hate speech. This requires a standardised conceptualisation of hate speech to ensure comparability of the study results. More fundamentally, social science is faced with the question of how to deal with the temptations of data access. Is it part of its role and mission to take on service tasks in return for data access and to carry out a kind of "third-party audit" (Meßmer and Degeling, 2023)? The attractiveness of data access for researchers could lead to research activities concentrating on the conclusively defined systemic risks arising from the design, functioning, use, or misuse of VLOPs and VLOSEs, with the result that other issues may be neglected.

Although Recital 5 of the DSA addresses the problem of the "intermediation and spread of unlawful or otherwise *harmful* information and activities", it is important to emphasise that legal regulation is limited to *illegal* content and should therefore not (and this cannot be ruled out) target legal hate speech, incivility, or a negative quality of discourse (Cornils, 2020). This means that hate speech must be countered in a differentiated way, depending on its intensity. Clearly illegal hate speech that is directly and immediately harmful (e.g., by inciting violence) should be removed by the platforms as quickly as possible in order to avoid contagion effects on third parties. Civil society actors or platform providers are called upon to address issues of discourse quality. Non-profit initiatives (e.g., having funding stabilised) should be strengthened in their commitment to more discursive diversity or the protection of the personal rights of those affected by hate speech (de Streel et al, 2020).

Platform providers, in turn, could label problematic, but not illegal, hate speech (e.g., negative stereotyping) as such, as they often already do in connection with disinformation. On the one hand, labelling hate speech can

make those affected feel less isolated and that the views expressed form a minority position. On the other, it can strengthen the enforcement of social norms and more civil communication behaviour among observers of hate speech (Blackwell et al, 2017; Katsaros et al, 2021). Moreover, warnings or other interventions can (quite successfully) encourage users to think twice before sharing problematic content (Katsaros et al, 2021), thus mitigating impulsive reactions encouraged by platform logics. In addition, platforms could be required (at least by opt-in) not to measure the relevance of user-generated content based on affective interactions (Tucker et al, 2018). Instead, algorithmic logics could be guided by such values as rationality, civility, and diversity (Friess and Eilders, 2015). The DSA already stipulates that VLOPs and VLOSEs must provide their users with a recommender system which is not based on profiling (Art. 38). However, the definition and operationalisation of such values is challenging, volatile, and has already been criticised for being paternalistic. Similarly, it is challenging to distinguish between occasionally subtle illegal hate speech and legally permissible but harmful speech when dealing with large amounts of content. The context, tone, and intent of speech are all significant here. At the same time, it must be made clear that content moderation only takes effect after hate speech has already been produced. It does not address the underlying causes of hate speech, such as radicalisation, which often stems from perceived injustice, the formation of an outwardly delineated group identity, or the propagation of ideologies in closed groups and offline networks. Such a sense of injustice can be reinforced by one-sided information (van den Bos, 2020). It is also unlikely that removing or blocking illegal hate speech will change the attitudes of those who create and disseminate it in the first place. As the EC itself stated in the context of the Code of Conduct on countering illegal hate speech online, notice and action procedures and the removal of content can only help address the symptoms (European Commission, 2020c).

Last but not least, it is not possible to draw direct conclusions about the individual impact of hate speech from its prevalence in social networks. In between are the individual visibility of hate speech in social network feeds, the prerequisite of having to recognise hate speech, different attitudes, experiences, processing strategies, and other intervening variables. The same is true for the effectiveness of removing or countering hate speech. Researchers and policy-makers should also consider the extent to which the development of perceived and content-analytically measured hate speech and the registered offences in this context can be explained by the fact that:

a) hate speech is an inconsistently defined and operationalised term; b) the intensity of use of digital platforms is increasing; c) the use of certain terms and public discourses are becoming more or less taboo, and levels of awareness and understanding of hate speech are changing; d) criminal law enforcement is intensifying; and e) measurement methods are becoming more accurate. This article provides some food for thought on how some of these issues could be effectively handled. In order to provide reliable answers to questions such as these, legal and communication sciences should more closely combine their different strengths and collaborate more intensively in future.

Acknowledgements

The authors would like to thank two anonymous reviewers, Rita Gsenger and Marie-Therese Sekwenz for their valuable feedback throughout the revision and publication process.

References

Adelberg, P. (2022) 'Hassrede in sozialen Netzwerken – Reichweite und Grenzen der Pflichten und Rechte der Netzwerkbetreiber', *Kommunikation & Recht*, 25(1), pp. 19–25.

Albert, J. (2024). *Researcher access to platform data: Experts weigh in on the Delegated Act* [Online]. DSA Observatory. Available at: https://dsa-observatory.eu/2024/11/29/researcher-access-to-platform-data-experts-weigh-in-on-the-delegated-act/ (Accessed: 2 January 2025).

Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S. and Stringhini, G. (2021) 'Understanding the effect of deplatforming on social networks', *13th ACM Web Science Conference 2021*, pp. 187–195.

Andres, R. and Slivko, O. (2021) *Combating online hate speech: the impact of legislation on Twitter* [Discussion Paper]. Leibniz-Zentrum für Europäische Wirtschaftsforschung. [Online]. Available at: https://ftp.zew.de/pub/zew-docs/dp/dp21103.pdf (Accessed: 21 January 2025).

Appelman, N. (2023) *Disparate content moderation: mapping social justice organisations perspectives on unequal content moderation harms and the EU platform policy*. Institute for Information Law, University of Amsterdam [Online]. Available at: https://dsa-observatory.eu/2023/10/31/research-report-on-disparate-content-moderation/ (Accessed: 30 December 2024).

Arun, A., Chhatani, S., An, J., & Kumaraguru, P. (2024). X-posing Free Speech: Examining the Impact of Moderation Relaxation on Online Social Networks. *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, 201–211. https://doi.org/10.18653/v1/2024.woah-1.15

Banks, J. (2010) 'Regulating hate speech online', *International Review of Law, Computers & Technology*, 24(3), pp. 233–239.

Bayer, J. and Bárd, P. (2020) *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches* [Online]. Policy Department for Citizens' Rights and Constitutional Affairs. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf (Accessed: 21 January 2025).

Berberich, M. (2023) '§ 5 Sorgfaltspflichten, Moderationsverfahren und prozedurale Fairness' in Steinrötter, B. (ed.) *Europäische Plattformregulierung*. Nomos, pp. 126–174.

Berry, J.M. and Sobieraj, S. (2016) *The outrage industry: political opinion media and the new incivility*. Reprint ed. New York: Oxford University Press.

Blackwell, L., Dimond, J., Schoenebeck, S. and Lampe, C. (2017) 'Classification and its consequences for online harassment: design insights from HeartMob', *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), pp. 1–19.

Bodden, N., Holec, H.A., Hoß, B., Ziegele, M. and Wilms, L. K. (2023) 'Vom Netz genommen. Die Auswirkungen von Deplatforming auf die Online-Kommunikation der extremen Rechten auf Telegram am Beispiel der Identitären Bewegung', *Medien & Kommunikationswissenschaft*, 71(3–4), pp. 266–284. Available at: https://doi.org/10.5771/1615-634X-2023-3-4-266.

Brauneck, J. (2024) 'Das Verantwortungsbewusstsein der Plattformbetreiber im Digital Services Act', *Neue Zeitschrift für Verwaltungsrecht*, 43(6), pp. 377–384.

Brugger, W. (2003) 'The treatment of hate speech in German constitutional law (Part I)', *German Law Journal*, 4(1), pp. 1–22.

Bryson, J.J. (2023) 'Human experience and AI regulation: what European Union law brings to digital technology ethics', *Weizenbaum Journal of the Digital Society*, 3(3). Available at: https://doi.org/10.34669/WI.WJDS/3.3.8.

Buchheim, J. (2022) 'Der Kommissionsentwurf eines Digital Services Act – Regelungsinhalte, Regelungsansatz, Leerstellen und Konfliktpotential' in Spiecker, I.,

Buiten, M., Streel, A. and Peitz, M. (2020) 'Rethinking liability rules for online hosting platforms', *International Journal of Law and Information Technology*, 28, pp.139–166.

Cauffman, C. and Goanta, C. (2021) 'A new order: the Digital Services Act and consumer protection', *European Journal of Risk Regulation*, 12(4), pp. 758–774.

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), pp. 114–212.

Cioffi, J.W., Kenney, M.F. and Zysman, J. (2022) 'Platform Power and Regulatory Politics: Polanyi for the Twenty-First Century' *New Political Economy*, 27(5), pp. 820–36.

'Charter of the Fundamental Rights of the Euroepan Union (2000/C 364/01)' (2000) *Official Journal of the European Communities* C 364/1, 18 December [Online]. Available at: https://www.europarl.europa.eu/charter/pdf/text_en.pdf (Accessed: 20 January 2025).

Coalition for Independent Technology Research (2024) *Blocking our right to know: surveying the impact of Meta's CrowdTangle shutdown* [Online]. Available at: https://independenttechresearch.org/wp-content/uploads/2024/07/CrowdTangle-Survey-Report-Final.pdf (Accessed: 20 January 2025).

Cohen-Almagor, R. (2011) 'Fighting hate and bigotry on the internet', *Policy & Internet*, 3(3), pp. 1–26.

Cole, M. D., Ukrow, J. and Etteldorf, C. (2020) *Zur Kompetenzverteilung zwischen der Europäischen Union und den Mitgliedstaaten im Mediensektor Eine Untersuchung unter besonderer Berücksichtigung medienvielfaltsbezogener Maßnahmen* [Online]. Institut für Europäisches Medienrecht. Available at: https://www.rlp.de/fileadmin/rlp-stk/pdf-Dateien/Medienpolitik/EMR_Gutachten_Zur_Kompetenzverteilung_im_Mediensektor.pdf (Accessed: 20 January 2025).

'COMMISSION RECOMMENDATION (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online', *Official Journal* L 63, 6 April. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0334 (Accessed 19 January 2025).

Cooper, H. (2018). 'Angela Merkel signals potential changes to online hate speech law', *Politico* 03 February [online]. Available at: https://www.politico.eu/article/angela-merkel-signals-potential-changes-to-germany-online-hate-speech-law/ (Accessed: 19 January 2025).

Cornils, M. (2020) *Designing platform governance: A normative perspective on needs, strategies, and tools to regulate intermediaries* [Online]. AlgorithmWatch. Available at: https://algorithmwatch.org/de/wp-content/uploads/2020/05/Governing-Platforms-legal-study-Cornils-May-2020-AlgorithmWatch.pdf (Accessed: 20 January 2025).

Courchesne, L., Ilhardt, J. and Shapiro, J. N. (2021) 'Review of social science research on the impact of countermeasures against influence operations', *Harvard Kennedy School Misinformation Review*, 13 September [Online]. Available at: https://doi.org/10.37016/mr-2020-79 (Accessed: 20 January 2025).

Das NETTZ, Gesellschaft für Medienpädagogik und Kommunikationskultur, HateAid and Neue deutsche Medienmacher*innen (eds.) (2024) *Lauter Hass - leiser Rückzug: Wie Hass im Netz den demokratischen Diskurs bedroht* [Online]. Kompetenznetzwerk Hass im Netz. Available at: https://kompetenznetzwerk-hass-im-netz.de/wp-content/uploads/2024/02/Studie_Lauter-Hass-leiser-Rueckzug.pdf (Accessed: 20 January 2025).

'Delegated Regulation (EU) 2024/436 of 20 October 2023 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines' (2024), *Official Journal* L [Online]. Available at: http://data.europa.eu/eli/reg_del/2024/436/oj (Accessed: 21 January 2025).

Democracy Reporting International (2024) *Access granted: why the European Commission should issue guidance on access to publicly available data now* [Online]. 9 September. Available at: http://democracy-reporting.org/en/office/global/publications/access-granted-why-the-european-commission-should-issue-guidance-on-access-to-publicly-available-data-now (Accessed: 30 December 2024).

De Streel, A., Defreyne, E., Jacquemin, H., Ledger, M., Michel, A., Innesti, A., Goubet, M. and Ustowski, D. (2020) *Online platforms' moderation of illegal content online. Law, practices and options for reform* [Online, study requested by the IMCO committee]. Policy Department for Economic, Scientific and Quality of Life Policies. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf (Accessed: 20 January 2025).

'Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')' (2000) *Official Journal* L 178, 17 July, pp. 1–16. Available at: http://data.europa.eu/eli/dir/2000/31/oj (Accessed: 19 January 2025).

'Directive (EU) 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive)' (2010) *Official Journal* L 303, 15 April, p. 69-92. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32010L0013.

Döhmann, G., Westland, M. and Campos, R. (eds.) *Demokratie und Öffentlichkeit im 21. Jahrhundert – zur Macht des Digitalen*. Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG, pp. 249–272.

Duffy, B. E. and Meisner, C. (2023) 'Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility', *Media, Culture & Society*, 45(2), pp. 285–304.

EPP Group. (2021) *Social media cannot be a lawless place* [Online]. Available at: https://www.eppgroup.eu/newsroom/social-media-cannot-be-a-lawless-place (Accessed: 17 January 2025).

Erjavec, K. and Kovačič, M. P. (2012) '"You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments', *Mass Communication and Society*, 15(6), pp. 899–920 [Online]. Available at: https://doi.org/10.1080/15205436.2011.619679 (Accessed 19 January 2025).

European Commission (2014a) REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law /* COM/2014/027 final */ [Online]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52014DC0027 (Accessed: 19 January 2025).

European Commission (2016) Code of Conduct on Countering illegal hate speech online [Online]. Available at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (Accessed: 19 January 2025).

European Commission (2020a) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the digital sector (Digital Markets Act) COM/2020/842 final [Online]. Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A842%3AFIN (Accessed: 19 January 2025).

European Commission (2020b) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC COM/2020/825 final [Online]. Available at: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN (Accessed: 20 January 2025).

European Commission (2020c) The Code of conduct on countering illegal hate speech online [Online]. Available at: https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135 / (Accessed: 19 January 2025).

European Commission (2021) C*ommunication from the Commission to the European Parliament and the Council. A more inclusive and protective Europe: Extending the list of EU crimes to hate speech and hate crime* [Online]. Available at: https://commission.europa.eu/document/download/926b3cb2-f027-40b6-ac7b-2c198a164c94_en?filename=COM_2024_146_1_EN.pdf (Accessed: 20 January 2025).

European Commission (2023a) *Application of the risk management framework to Russian disinformation campaigns* [Online]. Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2759/764631 (Accessed: 20 January 2025).

European Commission (2023b) *Commission opens formal proceedings against X under the Digital Services Act* [Online]. Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709 (Accessed: 21 January 2025).

European Commission (2024b) *Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act* [Online]. Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373 (Accessed: 21 January 2025).

European Commission (2024c) *Commission sends requests for information to 17 Very Large Online Platforms and Search Engines under the Digital Services Act* [Online]. Available at: https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-17-very-large-online-platforms-and-search-engines-under (Accessed: 21 January 2025).

European Commission (2024d) *Commission opens formal proceedings against TikTok under the Digital Services Act* [Online]. Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926 (Accessed: 21 January 2025).

European Commission (2025a) *CODE OF CONDUCT ON COUNTERING ILLEGAL HATE SPEECH ONLINE +* [Online]. Available at: https://ec.europa.eu/newsroom/dae/redirection/document/111777 (Accessed: 18 March 2025).

European Commission (2025b) *Supervision of the designated very large online platforms and search engines under DSA* [Online]. Available at: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses (Accessed: 18 March 2025).

European Commission (2025c) *The Code of conduct on countering illegal hate speech online +* [Online]. Available at: https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online (Accessed: 18 March 2025).

European Parliament (2024) *Briefing. Hate speech and hate crime must become crimes under EU law* [Online]. Available at: https://www.europarl.europa.eu/news/en/agenda/briefing/2024-01-15/11/hate-speech-and-hate-crime-must-become-crimes-under-eu-law (Accessed: 21 January 2023).

Fielitz, M. and Marcks, H. (2019) *Digital fascism: challenges for the open society in times of social media* [Online]. Berkeley Center for Right-Wing Studies Working Paper Series. Berkeley. Available at: https://escholarship.org/uc/item/87w5c5gp (Accessed: 20 January 2025).

Fielitz, M. and Schwarz, K. (2020) *Hate not found?! Das Deplatforming der extremen Rechten und seine Folgen* [Online]. Institut für Demokratie und Zivilgesellschaft. Available at: https://www.idz-jena.de/fileadmin//user_upload/Hate_not_found/WEB_IDZ_FB_Hate_not_Found.pdf (Accessed: 20 January 2025).

Flew, T., Martin, F. and Suzor, N. (2019) 'Internet regulation as media policy: rethinking the question of digital communication platform governance', *Journal of Digital Media & Policy*, 10(1), pp. 33–50.

'Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law' (2008) *Official Journal* L 328/55, 6 December [Online]. Available at: https://db.eurocrim.org/db/en/doc/1044.pdf (Accessed: 19 January 2025).

Friess, D. and Eilders, C. (2015) 'A Systematic Review of Online Deliberation Research' *Policy & Internet*, 7, pp. 319-339.

Fuchs, C. (2022) *Digital fascism*. Abingdon: Routledge.

Geese, A. (2022) *Europe Calling "DSA Deal: A constitution for the internet!"* [Online video]. 29 April. Available at: https://en.alexandrageese.eu/video/europe-calling-dsa-deal/ (Accessed: 19 January 2025).

Gelber, K. and McNamara, L. (2016) 'Evidencing the harms of hate speech', *Social Identities*, 22(3), pp. 324–341.

Gerdemann, S. and Spindler, G. (2023) 'Das Gesetz über digitale Dienste (Digital Services Act) (Teil 1)', *Gewerblicher Rechtsschutz und Urheberrecht*, 125(1–2), pp. 3–11.

'Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG)' *BGBl.* I 2017, p. 3351 [Online]. Available at: https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html (Accessed on: 20 January 2025).

Google and Youtube (2019) *Stellungnahme im Rahmen der öffentlichen Anhörung des Ausschusses für Recht und Verbraucherschutz des Deutschen Bundestages 15. Mai 2019* [Online]. Available at: https://kripoz.de/wp-content/uploads/2019/05/stellungnahme-frank-netzdg.pdf (Accessed on: 21 January 2025).

Gorwa, R., Binns, R. and Katzenbach, C. (2020) 'Algorithmic content moderation: technical and political challenges in the automation of platform governance', *Big Data & Society*, 7(1) [Online]. Available at: https://doi.org/10.1177/2053951719897945 (Accessed: 20 January 2025).

Griffin, R. (2023) 'The law and political economy of online visibility. Technology and regulation', *Technology and Regulation*, pp. 69–79.

Haimson, O. L., Delmonaco, D., Nie, P. and Wegner, A. (2021) 'Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: marginalization and moderation gray areas', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 466, pp. 1–35.

Hammer, D., Matlach, P., Gerster, L. and Baaken, T. (2021) *Fluchtwege. Wie das Netzwerkdurchsetzungsgesetz auf etablierten sozialen Medien durch die Verlinkung zu alternativen Plattformen umgangen wird* [Online]. Institute for Strategic Dialogue. Available at: https://www.isdglobal.org/wp-content/uploads/2021/08/Fluchtwege_050821_V4.pdf (Accessed: 20 January 2025).

Helberger, N. (2020) 'The political power of platforms: how current attempts to regulate misinformation amplify opinion power', *Digital Journalism*, 8(6), pp. 842–854.

Heldt, A. (2019) 'Let's meet halfway: sharing new responsibilities in a digital age,' *Journal of Information Policy*, 9, pp. 336–369.

Hestermann, T., Hoven, E. and Autenrieth, M. (2021) '"Eine Bombe, und alles ist wieder in Ordnung": Eine Analyse von Hasskommentaren auf den Facebook-Seiten reichweitenstarker deutscher Medien', *Kriminalpolitische Zeitschrift*, 4, pp. 204–214.

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., & Burghardt, K. (2023). Auditing Elon Musk's Impact on Hate Speech and Bots. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 1133–1137. https://doi.org/10.1609/icwsm.v17i1.22222

Hofmann, F. (2023a) 'Vor Art. 4 ff' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 111–139.

Hofmann, F. (2023b) 'Art. 7 Freiwillige Untersuchungen auf Eigeninitiative und Einhaltung der Rechtsvorschriften' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 175–183.

Hofmann, F. (2023c) 'Art. 8 Keine allgemeine Verpflichtung zur Überwachung oder aktiven Nachforschung' in. Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 183–191.

Hofmann, F. (2023d) 'Art. 9 Anordnungen zum Vorgehen gegen rechtswidrige Inhalte' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 191–206.

Hofmann, F. and Raue, B. (2023) 'Einleitung' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 31–48.

Holznagel, D. (2021) 'Chapter II des Vorschlags der EU-Kommission für einen Digital Services Act—Versteckte Weichenstellungen und ausstehende Reparaturen bei den Regelungen zu Privilegierung, Haftung & Herkunftslandprinzip für Provider und Online-Plattformen', *Computer und Recht*, 37(2), pp. 123–132.

Hong, M. (2022) 'Regulating hate speech and disinformation online while protecting freedom of speech as an equal and positive right – comparing Germany, Europe and the United States', *Journal of Media Law*, 14(1), pp. 76–96.

Jaursch, J. (2021) *Der DSA-Entwurf: Ehrgeizige Regeln, schwache Durchsetzungsmechanismen. Warum eine europäische Plattformaufsicht sinnvoll ist* [Online]. Stiftung Neue Verantwortung. Available at: https://www.stiftung-nv.de/sites/default/files/snv_dsa-aufsicht.pdf (Accessed: 20 January 2025).

Jaursch, J. and Lorenz-Spreen, P. (2024) *Researcher access to platform data under the DSA: questions and answers* [Online]. Available at: https://reclaimingautonomyonline.notion.site/Researcher-access-to-platform-data-under-the-DSA-Questions-and-answers-8f7390f3ae6b4aa7ad53d53158ed257c (Accessed: 30 December 2024).

Kalbhenn, J. C. and Hemmert-Halswick, M. (2021) 'EU-weite Vorgaben für die Content-Moderation in sozialen Netzwerken Kommentar zu dem Entwurf der Europäischen Kommission zu einem Digital Services Act', *Zeitschrift für Urheber- und Medienrecht*, 3, pp. 184–194.

Kapusta, I. (2024) 'Plattformregulierung 2.0: Die (un-)mittelbare Grundrechtsbindung Privater im Digital Services Act', in Laimer, S., Mittwoch, A.-C., Müller, T. and Staffler, L. (eds.) *Daten, Plattfomen, Smart Contracts*. Baden-Baden: Nomos, pp. 271–327.

Katsaros, M., Kim, J. and Tyler, T. (2024) 'Online Content Moderation: Does Justice Need a Human Face?' *International Journal of Human–Computer Interaction*, 40 (1), pp. 66–77.

Keipi, T., Näsi, M., Oksanen, A. and Räsänen, P. (2017*) Online hate and harmful content: cross-national perspectives*. Abingdon: Routledge.

Kettemann, M. C. (2019) *Stellungnahme als Sachverständiger für die öffentliche Anhörung zum Netzwerkdurchsetzungsgesetz auf Einladung des Ausschusses für Recht und Verbraucherschutz des Deutschen Bundestags* [Online]. Leibniz-Institut für Medienforschung| Hans-Bredow-Institut. Available at: https://kripoz.de/wp-content/uploads/2019/05/stellungnahme-kettemann-netzdg.pdf (Accessed: 19 January 2025).

King, G. and Persily, N. (2019) A new model for industry-academic partnerships. *PS: Political Science and Politics*, 53(4), pp. 703–709.

Klinger, U. and Ohme, J. (2023) *What the scientific community needs from data access under Art. 40 DSA: 20 points on infrastructures, participation, transparency, and funding* [Online]. Available at: https://doi.org/10.34669/WI.WPP/8.2 (Accessed: 19 January 2025).

Klonick, K. (2018) The new governors: the people, rules, and processes governing online speech. *Harvard Law Review*, 131, pp. 1598–1670.

Koehler, M. (2024) 'Artikel 7 Freiwillige Untersuchungen' in Mueller-Terpitz, R. and Koehler, M. (eds.) *Digital Services Act*. München: C.H. Beck, pp. 106–118.

Kommunikationsplattformen-Gesetz, BGBl. I Nr. 151/2020 [Online]. Available at: https://www.ris.bka.gv.at/eli/bgbl/I/2020/151/20201223 (Accessed: 19 January 2025).

Kohl, U. (2022) 'Platform regulation of hate speech – a transatlantic speech compromise?' *Journal of Media Law*, 14(1), pp. 25-49.

Koreng, A. (2017) 'Hate-Speech im Internet: Eine rechtliche Annäherung', *Kriminalpolitische Zeitschrift*, 3, pp. 151–159.

Kuczerawy, A. (2021) 'The good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act', *Verfassungsblog*, 12 January [Online]. Available at: https://verfassungsblog.de/good-samaritan-dsa/ (Accessed: 30 December 2024).

Kupferschmidt, K. (2023) 'Twitter's plan to cut off free data access evokes 'fair amount of panic' among scientists', *Science*, 8 February [Online]. Available at: https://www.science.org/content/article/twitters-plan-cut-free-data-access-evokes-fair-amount-panic-among-scientists (Accessed: 19 January 2025).

Landesanstalt für Medien NRW (2023) *Hate Speech. Forsa-Studie 2023. Zentrale Untersuchungsergebnisse* [Online]. Available at: https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2023_Praesentation.pdf (Accessed: 20 January 2025).

Latzer, M., Saurwein, F. and Just, N. (2019) ‚Assessing Policy II: Governance-Choice Method' in Van Den Bulck, H., Puppis, M., Donders, K. and Van Audenhove, L. (eds.) *The Palgrave Handbook of Methods for Media Policy Research*. Cham: Springer International Publishing, pp. 557-574.

Legner, S. (2024) 'Der Digital Services Act - Ein neuer Grundstein der Digitalregulierung', *Zeitschrift für Urheber- und Medienrecht*, 68(2), pp. 99–111.

Lee-Won, R.J., White, T.N., Song, H., Lee, J.Y. and Smith, M.R. (2020) 'Source magnification of cyberhate: affective and cognitive effects of multiple-source hate messages on target group members', *Media Psychology*, 23(5), pp. 603–624.

Liesching, M. (2021) *Das NetzDG in der praktischen Anwendung: Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes*. Carl Grossmann [Online]. Available at: https://doi.org/10.24921/2021.94115953 (Accessed: 20 January 2025).

LOI n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (1). Journal Officiel de la République Française n°0156, p.11, 25 June [Online]. Available at: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000042031970 (Accessed: 19 January 2025).

Mchangama, J. and Fiss, J. (2019) *The digital Berlin Wall: how Germany (accidentally) created a prototype for global online censorship* [Online]. Justitia. Available at: http://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf (Accessed: 20 January 2025).

Meßmer, A.-K. and Degeling, M. (2023) *Auditing recommender systems* [Online]. Stiftung Neue Verantwortung. Available at: https://www.stiftung-nv.de/de/publication/auditing-recommender-systems (Accessed: 30 December 2024).

Meta (2024a) *CrowdTangle*. [Online]. Available at: https://transparency.meta.com/de-de/researchtools/other-datasets/crowdtangle/ (Accessed: 30 December 2024).

Meta (2024b) *Meta Content Library and API* [Online]. Available at: https://transparency.meta.com/en-gb/researchtools/meta-content-library/ (Accessed: 21 January 2025).

Müller, K. and Schwarz, C. (2021) 'Fanning the flames of hate: social media and hate crime', *Journal of the European Economic Association*, 19(4), pp. 2131–2167.

Newman, N. (2023) 'Executive summary and key findings' in Newman, N., Fletcher, R., Eddy, K., Robertson, C.T. and Nielsen, R.K. (eds.) *Reuters Institute Digital News Report 2023*. Oxford: Reuters Institute for the Study of Journalism, pp. 9–29.

Paasch-Colberg, S., Trebbe, J., Strippel, C. and Emmer, M. (2022) 'Insults, criminalisation, and calls for violence: forms of hate speech and offensive language in German user comments on immigration', in Monnier, A., Boursier, A. and Seoane, A. (eds.) *Cyberhate in the Context of Migrations*. Cham: Springer International Publishing, pp. 137–163.

Pohlmann, J., Barbaresi, A. and Leinen, P. (2023) 'Platform regulation and "overblocking" – the NetzDG discourse in Germany', *Communications*, 48(3), pp. 395–419.

Price, L. (2021) 'Platform responsibility for online harms: towards a duty of care for online hazards', *Journal of Media Law*, 13(2), pp. 238–261.

Ranaivoson, H. and Domazetovikj, N. (2023) 'Platforms and exposure diversity: towards a framework to assess policies to promote exposure diversity', *Media and Communication*, 11(2), pp. 379-391.

Raue, B. (2023a) 'Art. 16 Melde- und Abhilfeverfahren' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 285–313.

Raue, B. (2023b) 'Art. 20 Internes Beschwerdemanangementsystem' in Hofmann, F. and Raue, B. (eds.) *Digital Services Act*. Baden-Baden: Nomos, pp. 341–360.

Recuero, R. (2024) 'The platformization of violence: toward a concept of discursive toxicity on social media', *Social Media + Society*, 10(1), [Online]. Available at: https://doi.org/10.1177/20563051231224264 (Accessed: 20 January 2025).

'Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online' (2021) Official Journal L172/79, 17 May, [Online]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32021R0784 (Accessed: 21 January 2025).

'Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)' (2022) *Official Journal* L 265, 12 October, pp. 1-66 [Online]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R1925 (Accessed: 19 January 2025).

'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)' (2022) *Official Journal* L 277, 27 October, pp. 1-102. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065 (Accessed: 19 January 2025).

Rieder, B. and Hofmann, J. (2020) 'Towards platform observability', *Internet Policy Review*, 9(4), pp. 1–28.

Rogers, R. (2020) 'Deplatforming: following extreme internet celebrities to Telegram and alternative social media', *European Journal of Communication*, 35(3), pp. 213–229.

Rüdiger, T.-G. (2019) 'Polizei im digitalen Raum', *Aus Politik und Zeitgeschichte*, 69(21–23), pp. 18–23.

Reporters Without Borders (2017) *Russian bill is copy-and-paste of Germany's hate speech law* [Online]. 19 July. Available at: https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law (Accessed: 30 December 2024).

Ruschemeier, H. (2024). Flagging trusted flaggers. *Verfassungsblog*, 4 November [Online]. Available at: https://doi.org/10.59704/6c2c9f4cc624f31a (Accessed: 30 December 2024).

Scheer, O., Vela, J. H., & Jahn, T. (2025, January 24). EU-Kommission: Untersuchung zu X abgeschlossen – Musk droht Millionenstrafe. *Handelsblatt*. https://www.handelsblatt.com/politik/international/eu-kommission-untersuchung-zu-x-abgeschlossen-musk-droht-millionenstrafe/100102819.html

Schulz, W. (2019) 'Regulating intermediaries to protect privacy online – the case of the German NetzDG' in Schulz, W., Kettemann, M.C., and Heldt., A.P. (eds.) *Probleme und Potenziale des NetzDG ein Reader mit fünf HBI-Expertisen* [*Problems and potentials of the NetzDG*]. Hamburg: Verlag Hans-Bredow-Institut, pp. 7–19.

Seiling, L., Ohme, J., & Klinger, U. (2024). *Response to the consultation on the delegated regulation on data access provided for in the Digital Services Act*. Weizenbaum Institute [Online]. Available at: https://www.weizenbaum-institut.de/media/Publikationen/Weizenbaum_Policy_Paper/Weizenbaum_Policy_Paper_11.pdf (Accessed: 20 January 2025).

Senftleben, M. (2024) 'Human rights outsourcing and reliance on user activism in the DSA', *Verfassungsblog*, 21 February [Online]. Available at: https://verfassungsblog.de/human-rights-outsourcing-and-reliance-on-user-activism-in-the-dsa/ (Accessed: 30 December 2024).

Siegel, A. A. (2020) 'Online hate speech', in Persily, N. and Tucker, J.A. (eds.) *Social media and democracy: the state of the field, prospects for reform*. Cambridge: Cambridge University Press, pp. 56–88.

Specht-Riemenschneider, L. (2021) *Studie zur Regulierung eines privilegierten Zugangs zu Daten für Wissenschaft und Forschung durch die regulatorische Verankerung von Forschungsklauseln in den Sektoren Gesundheit, Online-Wirtschaft, Energie und Mobilität (Studie im Auftrag des Bundesministeriums für Bildung und Forschung)*. Fachbereich Rechtswissenschaft der Universität Bonn [Online]. Available at: https://www.jura.uni-bonn.de/fileadmin/Fachbereich_Rechtswissenschaft/Einrichtungen/Lehrstuehle/Specht/Dateien/2021-08-25-LSR.pdf (Accessed: 20 January 2025).

Spindler, G. (2017) 'Der Regierungsentwurf zum Netzwerkdurchsetzungsgesetz - europarechtswidrig?' *Zeitschrift für Urheber- und Medienrecht*, 61(6), pp. 473–487.

Sponholz, L. (2023) ‚Hate speech' in Strippel, C., Paasch-Colberg, S., Emmer, M. and Trebbe, J. (eds.) *Challenges and Perspectives of Hate Speech Research.* Digital Communication Research Vol. 12. Berlin: Böhland & Schremmer, pp. 143-163.

Stark, B., Stegmann, D. and Jürgens, P. (2020) *Are algorithms a threat to democracy? The rise of intermediaries: a challenge for public discourse*. Algorithm Watch [Online]. Available at: https://algorithmwatch.org/en/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch.pdf (Accessed: 20 January 2025).

Stippler, F., Scheuer, S., Kort, K., Holtermann, F., & Soares, P. A. de S. (2025, January 8). Tech-Konzern: Meta beendet Faktenchecks auf Facebook und Instagram. *Handelsblatt*. https://www.handelsblatt.com/technik/it-internet/tech-konzern-meta-beendet-faktenchecks-auf-facebook-und-instagram/100099044.html

The Economist. (2018) 'In Germany, online hate speech has real-world consequences', *The Economist* 12 January [Online]. Available at: https://www.economist.com/graphic-detail/2018/01/12/in-germany-online-hate-speech-has-real-world-consequences (Accessed: 19 January 2025).

'Treaty on the Functioning of the European Union' (2012) *Official Journal* C 326, 26 October, pp. 47-390 [Online]. Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:12012E/TXT:en:PDF (Accessed: 21 January 2025).

Tucker, J. et al (2018) 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.' *SSRN Electronic Journal* [Online] Available at: https://doi.org/10.2139/ssrn.3144139 (Accessed: 21 January 2025).

Udupa, S., Gagliardone, I. and Hervik, P. (eds.) (2021) *Digital hate: the global conjuncture of extreme speech*. Bloomington: Indiana University Press.

Valerius, B. (2020) 'Hasskriminalität – Vergleichende Analyse unter Einschluss der deutschen Rechtslage', *Zeitschrift für die gesamte Strafrechtswissenschaft*, 132(3), pp. 666–689.

Van den Bos, K. (2020) 'Unfairness and radicalization', *Annual Review of Psychology*, 71(1), pp. 563–588.

Van Drunen, M. Z. and Noroozian, A. (2024) 'How to design data access for researchers: a legal and software development perspective', *Computer Law & Security Review*, 52 [Online].. Available at: https://doi.org/10.1016/j.clsr.2024.105946 (Accessed: 19 January 2025).

Wagner, E. (2019) *Intimisierte Öffentlichkeiten: Pöbeleien, Shitstorms und Emotionen auf Facebook*. Bielefeld: Transcript.

Williams, M. L., Burnap, P., Javed, A., Liu, H. and Ozalp, S. (2020) 'Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime', *The British Journal of Criminology*, 60(1), pp. 93–117.

'X Corp v. Center for countering digital hate Inc.' (2023) Case 3:23-cv-03836 [Online]. Available at: https://s3.documentcloud.org/documents/23892523/x-corp-v-center-for-countering-digital-hate.pdf (Accessed: 21 January 2025).