NEW SCHOOL SPEECH REGULATION AND ONLINE HATE SPEECH: A CASE STUDY OF GERMANY'S NETZDG

Rachel Griffin
PhD candidate in law
Sciences Po School of Research

Contents

Abbreviations

AfD	Alternative for Germany
AVMSD	Audiovisual Media Services Directive
BfJ	Federal Office of Justice
BKA	Federal Criminal Police
BMJV	Federal Ministry for Justice and Consumer Protection
CSO	civil society organisation
DSA	Digital Services Act
ECD	E-Commerce Directive
GÄNDG	Network Enforcement Act Amendment Act
GRH	Law Against Right-Wing Extremism and Hate Crime
NetzDG	Network Enforcement Act
NSSR	new school speech regulation
StGB	Criminal Code
TMG	Telemedia Act

Abstract

Germany's 2017 NetzDG law is an example of 'new school speech regulation' (Balkin, 2014), which restricts speech by coercing intermediaries into censoring users, rather than coercing speakers directly. It is the first such measure which specifically targets hate speech on social media, by requiring large platforms to operate complaints procedures which ensure illegal content is rapidly removed. Numerous other countries have since adopted similar regulations. This paper takes NetzDG as a case study to evaluate the effectiveness of this regulatory model.

A review of relevant empirical literature shows that many features of social media platforms actively promote hate speech. Key factors include algorithmic recommendations, which frequently promote hateful ideologies; social affordances which let users encourage or disseminate hate speech by others; anonymous, impersonal environments; and the absence of media 'gatekeepers'. In mandating faster content deletion, NetzDG only addresses the last of these, ignoring other relevant factors. Moreover, reliance on individual user complaints to trigger platforms' obligations means hate speech will often escape deletion. Interviews with relevant civil society organisations (CSOs) confirm these flaws of the NetzDG model. From their perspectives, NetzDG has had little impact on the prevalence or visibility of online hate speech, and its reporting mechanisms fail to help affected communities.

NetzDG represents an incremental, narrow approach to a complex sociotechnical problem which requires more fundamental regulatory reform. In this regard, it shows the limitations of censorship-based new school speech regulation. Rules prescribing censorship of narrowly-defined content categories are ill-suited to large-scale, networked, algorithmically-curated social media, where other governance mechanisms influence user behaviour more than content deletion. The paper advocates a more systemic and preventive regulatory approach. Platforms should be required to take public interest considerations into account in all design and governance processes, aiming to shape platform environments to actively discourage users from posting or viewing hate speech, rather than simply deleting it afterwards.

1. Introduction

In recent years, Germany has experienced rising far-right terrorism, electoral successes by the far-right AfD party, the discovery of extremist networks within the military and police, and record hate crime figures (Koehler, 2018; Connolly, 2020, 2021). Extremism and hateful ideologies pose serious threats. In Germany and elsewhere, hate speech and other harmful social media content have become a particular focus of academic and political debates.

Balkin (2014) observes that governments increasingly seek to address such issues through what he terms 'new school speech regulation' (NSSR). Scale, decentralisation, anonymity and international reach mean 'old-school regulation' of online communication – targeting speakers directly – is often infeasible. However, essential digital infrastructure is controlled by powerful private intermediaries – like social media companies – which governments can conscript into regulating users.

This paper investigates whether the predominant model of NSSR, which mandates deletion of unlawful content, effectively prevents hate speech on social media. Social media represent a distinctive media environment, raising distinct policy considerations. Empirical literature shows that platform features often actively promote hateful content and ideologies, or facilitate their dissemination. A large body of legal literature has analysed social media governance; however, this scholarship has focused particularly on free speech, transparency, and platform accountability. The effectiveness of different regulatory models in preventing hate speech has been comparatively under-explored.

To address this, I use Germany's 2017 *Netzwerkdurchsetzungsgesetz* (NetzDG) as a case study to evaluate the effectiveness of NSSR in this context. I conclude that NetzDG's 'old-fashioned' NSSR strategy, with legalistic rules mandating deletion of specified content categories, is profoundly unsuited to regulating contemporary social media, which influence user behaviour – including hate speech – through many mechanisms other than content deletion.

This contributes new insights to the literature on NetzDG. Compared to its much-debated free speech implications, surprisingly little research evaluates NetzDG's effectiveness against hate speech. One reason is the lack of accessible quantitative data (Tworek and Leerssen, 2019). This paper takes an alternative, qualitative approach. As well as drawing on existing empirical research to inform my analysis of NetzDG's regulatory strategy, I investigate how its effects are perceived

by CSOs representing affected communities. This paper also contributes to the wider literature on social media regulation. I offer some generalisable insights into the flaws of a globally-influential regulatory model, and recommendations which could inform future national and EU regulations.

The paper proceeds as follows. Section 2 reviews empirical literature on the causes and dynamics of hate speech on social media. Section 3 discusses legal scholarship on social media regulation, including Balkin's NSSR theory. Section 4 presents NetzDG as a case study of NSSR. Section 5 evaluates how effectively NetzDG addresses factors known to drive online hate speech, informed by my expert interviews. Finally, section 6 discusses key findings and offers policy recommendations. I argue that future regulations should – unlike NetzDG – take a systemic and preventive approach, considering the whole sociotechnical environment in which hate speech emerges.

2. Hate speech

a. Definitions

Definitions of hate speech vary widely. Reviewing definitions from several legal systems, Sellars (2016) identifies eight characteristic elements. These include causing or intending harm, and inciting discrimination or violence, but the most essential element is targeting someone based on membership in a protected group. These groups typically include core personal identity characteristics associated with discrimination, such as race, religion and gender. Hate speech is not a purely objective category, but is used rhetorically, to condemn alleged violations of social norms (Gagliardone, 2019; Udupa and Pohjonen, 2019). Post (2009) highlights that speech following elite civility norms is rarely considered hate speech, even if factually likely to encourage discrimination or violence. This does not mean the term, or the norms it enforces, are without value. However, it should be remembered when assessing policies like NetzDG that legal definitions of hate speech do not cover all discriminatory or hate-promoting speech.

In Germany, §130(1) of the criminal code (StGB) criminalises insults violating human dignity and incitement of hatred or violence based on nationality, race, religion or ethnicity. Notably, this excludes commonly-protected characteristics like gender and sexuality. However, §130(1) is best regarded as a subcategory of Germany's legal provision for hate speech. Other criminal speech acts, such as inciting or threatening a crime (§111 and §241 StGB), would also frequently fall within

Sellars' multifactorial definition, if they targeted protected groups. Hate speech is widely discussed in German media and politics in this broader sense (Gollatz and Jenner, 2018).

The *Bundeskriminalamt*'s (BKA) hate crime statistics use a broad definition[1]: crimes motivated by prejudice against nationality, ethnicity, skin colour, religion, social status, disability, gender, sexuality, or appearance (BKA, 2020b). For present purposes, any criminal speech act[2] meeting this motivation-based definition will be regarded as hate speech. This definition is broad enough to enable consideration of the social harms caused by hate speech; at the same time, by only including criminal speech, it aligns with the policy goals of NetzDG, which does not aim to comprehensively regulate any kind of harmful online discourse but only to strengthen the enforcement of specified criminal law provisions. As noted, not all harmful and discriminatory speech is criminal. However, from a policy perspective, non-criminal but discriminatory speech remains relevant as part of the social context in which hate speech occurs.

The vast majority of Germany's recorded hate crimes are racist (BKA, 2020a), and anti-migrant hate speech was a key motivation behind NetzDG. In academic literature, racist and far-right online hate speech appear particularly well-studied – probably because they often involve organised networks (Daniels, 2018; Lewis, 2019). Racist hate speech is therefore a particular focus of this paper. This should not be taken to downplay the presence or seriousness of hate speech against other groups. For example, female public figures frequently experience gender-based online harassment, including potentially criminal threats and privacy violations which could meet the above definition (Abé et al., 2021). Incidents recorded as hate crimes may not reflect the full spectrum of hate speech happening in Germany.

b. Hate speech on social media

i. Prevalence and consequences

Evidence on the overall prevalence of hate speech on social media is mixed. Based on random sampling, Facebook (2021b) estimates that 0.05% of viewed posts contain hate speech. Academic estimates are similarly low (Siegel, 2020). However, surveys in Germany and elsewhere find most users have encountered hate speech (Landesanstalt für Medien NRW, 2018; Reinemann et al.,

[1] This definition exists for statistical purposes and does not have legal force. However, it is useful in providing a list of identity characteristics considered relevant in this context in Germany.
[2] Appendix III provides a list of speech-based offences covered by NetzDG.

2019) – probably because it is often widely distributed (Kümpel and Rieger, 2019; Siegel, 2020). Facebook's views-based prevalence estimate takes distribution into account. However, 0.1-11% of a vast number – Facebook has 1.84 billion daily users (Facebook, 2021a) – is in any case substantial. Moreover, aggregated prevalence across a platform does not capture distribution of risk, which may disproportionately affect certain groups (Ananny, 2019).

Online hate speech harms victims emotionally and materially (Gerstenfeld, 2017). Although free expression and hate speech regulation are often framed as contradictory, hate speech limits free expression (Citron, 2014; douek, 2021): there is evidence that victims respond by withdrawing from discussions (Barnidge et al., 2019; Geschke et al., 2019; Stark and Stegmann, 2020) and/or self-censoring (Duguay et al., 2020). Women and minorities in public life (e.g. journalists, politicians) are particularly affected (Lamensch, 2021). Hate speech also affects non-targeted groups, increasing prejudices (Sorel et al., 2017). Anecdotally, many violent extremists have been influenced by online hate (Daniels, 2018). Some evidence suggests a more general link: Müller and Schwarz (2018) find an apparent causal association between Facebook access and anti-refugee violence in German municipalities.

ii. Causes and dynamics

Empirical literature shows that many platform features actively encourage hate speech. Algorithmic recommendations have attracted particular criticism. Many authors suggest that engagement-maximising algorithms systematically favour content provoking strong reactions, including hate and extremism (Vaidhyanathan, 2018; Whittaker et al., 2021; for a recent illustration see Hagey and Horwitz, 2021). There is some evidence for this claim. A leaked internal Facebook study from 2016 revealed that over one-third of Germany's large political groups promoted hate speech, and 64% of users joining were prompted by Facebook's recommendations (Horwitz and Seetharaman, 2020). German YouTube recommends progressively further-right videos to users watching right-wing content (Rauchfleisch and Kaiser, 2020).

However, Lewis (2020) warns against focusing exclusively on recommendations; other platform features and affordances also contribute. This has been documented in detailed studies on Reddit (Massanari, 2017), YouTube (Munger and Phillips, 2020), Instagram, Vine and Tinder (Duguay et al., 2020). Impersonal, sometimes anonymous environments reinforce in-group/out-group identities, encouraging prejudice (Keum and Miller, 2018; Stark and Stegmann, 2020), and may

encourage aggression by reducing inhibitions (Barak, 2005; Keum and Miller, 2018). Social media bypass traditional media gatekeepers, enabling extremists to present themselves on favourable terms (Klein, 2012; Rauchfleisch and Kaiser, 2020). On the 'demand side', without gatekeepers, prejudiced users can easily access hate content (Hosseinmardi et al., 2020); this demand creates and financially supports a supply (Munger and Phillips, 2020).

Social and interactive affordances are particularly significant (Matamoros-Fernández, 2017). They can facilitate the emergence of 'toxic technocultures' whose social norms encourage hate speech (Massanari, 2017). In combination with algorithmic recommendations, these affordances let users encourage and disseminate hate speech without themselves posting anything illegal (Ben-David and Matamoros-Fernández, 2016). For example, liking and commenting on posts lets users increase their visibility without posting themselves, since engagement-maximising algorithms typically promote posts which are attracting activity (Leerssen, 2020). Conversely, legal posts may deliberately encourage hateful comments. An ex-Facebook employee recently criticised the promotion of 'hate bait' posts from right-wing media, commenting that 'we reward them fantastically', as high engagement leads to algorithmic promotion (Mac and Silverman, 2020).

3. Regulating social media

a. New school speech regulation

Legal scholars have theorised extensively how platforms regulate users' speech and how states regulate this process. An influential contribution in the latter category is Balkin's (2014) theory of 'new school speech regulation' (NSSR), describing state regulation which coerces intermediaries into restricting speakers instead of coercing speakers directly.

Balkin first notes the importance of communications infrastructure for the meaningful exercise of free speech. Essential digital infrastructure is controlled by large corporations, whose power to control users' speech *en masse* using technological measures like filtering makes them attractive targets for regulation (Balkin, 2014). This is not unique to speech regulation, but reflects a broader trend of conscripting 'gatekeeper' companies as regulators (Van Loo, 2020). Balkin (2018a) describes a shift from dyadic to triadic regulation, with speakers, intermediaries and governments at the three points[3]. Although this triadic structure could involve many different regulatory

[3] As Balkin notes, this collapses some relevant distinctions; for an expanded version see Papaevangelou (2021).

interventions, Balkin's conceptual focus is clearly on blocking services and removing content (hence his discussion of 'censorship').

Balkin is particularly concerned with NSSR's implications for free speech. Alongside state co-option of private power, he identifies two concerning features. First, collateral censorship: if intermediaries are liable for users' speech (a common NSSR tactic), their incentive is to over-censor to minimise liability risks. Second, digital prior restraint: intermediaries often prevent speech *ex ante* rather than sanctioning it afterwards. NSSR can also operate through soft power, as when the US government encouraged WikiLeaks' cloud provider, domain name registry and payment processors to block them (Balkin, 2014). Such extra-legal regulation raises particular free speech concerns (Leerssen, 2015).

Balkin (2014) simply lists social media alongside other intermediaries. In later work, he addresses their distinctive regulatory considerations (Balkin, 2016, 2018a, 2018c, 2020). However, he focuses primarily on restraining private power, rather than on implementing state speech regulation. Balkin does not argue that NSSR is never justified, or that platforms should not ban hate speech. Indeed, as US law prevents the government from doing this (Sellars, 2016), he considers its necessity an argument against publicly-provided social media (Balkin, 2020). However, Balkin has not examined in detail how – where NSSR targeting social media is justified – it can be implemented effectively.

b. Limitations of NSSR

As noted above, Balkin's account of NSSR focuses primarily on regulations mandating censorship. Examples of NSSR targeting social media, like NetzDG, the French *loi Avia* and the Austrian *Kommunikationsplattformengesetz*, also take a censorship-focused approach, by mandating faster and more comprehensive moderation (deletion) of illegal content.

Empirical literature indicates that moderation helps reduce hate speech – up to a point. Hate speech is facilitated by bypassing traditional media gatekeepers (Klein, 2012; Hosseinmardi et al., 2020); deleting hate content essentially reintroduces gatekeeping. Affected users can switch accounts or platforms, but not all do so, and those who do lose followers and visibility (Berger and Perez, 2016; Fielitz et al., 2020). Users determined to find or share hate speech can generally manage to. However, barriers to access deprive extremists of 'two key resources: reach and attention' (Fielitz

et al., 2020, p54), making it harder to organise, access financial resources, and reach beyond existing supporters.

However, NSSR mandating moderation has three major limitations. First, this evidence in support of content moderation examines its effects on users and content which are actually banned – but only a portion of hate speech is[4]. Current moderation practices rely on a combination of automated content recognition and user reporting. Neither is reliable enough to be an adequate solution; both produce frequent false negatives (overlooking hate speech) and false positives (removing non-hateful content). Much hate speech goes unreported because it is seen by sympathetic audiences, or because platform affordances make reporting laborious (Crawford and Gillespie, 2016; Duguay et al., 2020). Reporting has also been used maliciously against victims of discrimination (Matamoros-Fernández, 2017; Duguay et al., 2020). Automated moderation is notoriously unreliable, particularly for complex, context-dependent categories like hate speech (Gorwa et al., 2020; Laaksonen et al., 2020). Widely-available commercial moderation software disproportionately removes speech from marginalised groups (Kayser-Bril, 2020; see also Duarte et al., 2017). Although content moderation is useful, its vulnerability to misuse and bias and its failure to catch much hate speech make it a flawed, partial solution.

Second, even if moderation were more reliable, it is not the primary and certainly not the only factor shaping user interactions. As section 2 showed, features like recommendations, interactive affordances and user cultures may make users feel comfortable behaving aggressively, prompt them to view hate speech, help them disseminate hate speech by others, and/or enable creators to profit from hateful messages. Simply removing some obvious hate content will not address these issues. These features also distinguish social media companies from the other intermediaries Balkin discusses. Unlike, say, payment processors deciding whether to serve WikiLeaks, social media platforms make complex decisions which do not reduce to binary choices between censorship and non-intervention (douek, 2021).

Finally, if the aim is not just to condemn but to prevent criminal hate speech, measures should not focus exclusively on criminal content: they must also consider the broader ideologies and norms being promoted. Legal content may strategically encourage hate speech, or increase its visibility. Yet state-mandated censorship of 'harmful but legal' content raises major free speech and rule-of-

[4] Ex-employees have estimated that Facebook deletes around 5% of hate speech (Mac and Silverman, 2020). Leaked internal documents from Facebook also provide estimates in the low single figures (Seetharaman et al., 2021).

law concerns (Harbinja et al., 2019). There are therefore normative as well as practical reasons not to rely solely on moderation. Discouraging or de-amplifying harmful content – for example, ensuring it is not recommended, or designing affordances to discourage aggressive behaviour – can be simultaneously more effective, and less restrictive of free speech (Heldt, 2019a; Bowers and Zittrain, 2020; douek, 2021).

c. Alternative approaches

In light of such considerations, scholars have considered regulatory interventions beyond censorship. For example, douek (2021) argues platforms' regulation of users' speech should be guided by proportionality (balancing free speech and other interests) and probability (focusing on large-scale, systemic balancing, rather than individual cases). Proportionality implies 'remedial flexibility', utilising non-censorship interventions like labelling, demonetisation and restricting sharing (pp26-27)[5]. State regulation should 'institutionalize, incentivize and verify the systemic balancing of platforms', e.g. through audits (p64).

Other scholars examine how platforms regulate users through recommendations, affordances and design, exercising 'opinion power' (Neuberger, 2018; Helberger, 2020) or 'organisational control' (Van Drunen, 2020) to shape interactions, information flows, and public opinion. Like Balkin, these scholars focus primarily on restraining private abuses of power. Helberger (2020) argues NetzDG is 'potentially counterproductive and dangerous for democracy' because it strengthens platforms' unaccountable opinion power (p845). She argues regulation should strengthen transparency requirements and 'countervailing powers' like traditional media and civil society (p848). This would certainly be positive for media democracy generally, but it is unclear how it would achieve NetzDG's original aim of preventing hate speech. Moreover, opinion power is not inherently undesirable: platforms selecting and organising information cannot be neutral or value-free (Cobbe and Singh, 2019). Often platforms' opinion power appears to actively encourage hate speech – but regulations could also incentivise the opposite.

Platform design choices are a form of regulation, constraining and influencing how users communicate (Lessig, 1999; Gorwa, 2019b). Since empirical research indicates that platforms' algorithms and other design choices significantly influence hate speech, they would seem promising areas for intervention. Citron (2014) suggests design and visual cues could counter the disinhibiting

[5] Goldman (2021) identifies 36 types of intervention.

effects of anonymity, while Hartzog and Selinger (2015) argue that technical affordances can discourage harassment by increasing its transaction costs: for example, making it harder to search for victims' profiles. Supporting these arguments, Copland (2020) shows that 'quarantining' problematic Reddit forums (meaning they were not banned, but made more difficult to access) successfully reduced hate posts. Other platforms have tested design interventions. Facebook and YouTube claim to algorithmically demote 'borderline' content (YouTube, 2019; Rosen, 2020). Twitter and YouTube have tested behavioural prompts (another of Hartzog and Selinger's suggestions): where automated classification flags a draft comment, users can still post it, but are informed it may be offensive and invited to reconsider (YouTube, 2020; Butler and Parella, 2021). Based on platforms' own published results, such interventions seem promising. However, we lack independent research investigating their impact.

For interventions to be tested and implemented systematically, rather than as ad hoc voluntary measures, some form of legal regulation is necessary. Suzor et al. (2019) suggest that the non-binding UN Guiding Principles on Human Rights ground a legal responsibility for platforms to mitigate abuse and misogyny, including through design, but acknowledge that binding regulation is necessary to operationalise this responsibility. Cobbe and Singh (2019) have put forward one such proposal, arguing that platforms should be banned from operating recommendation systems for user-generated content unless they do so 'responsibly' and avoid recommending illegal content. This is helpful in proposing concrete regulatory solutions beyond moderation, although it does not address other design features, or provide much detail on how this general duty of responsibility could be operationalised.

4. NetzDG: a case study

a. Case selection and methodology

NetzDG was passed in 2017, and threatens large platforms with heavy fines for systematic failures to promptly remove illegal content after receiving complaints. This is a classic example of censorship-based NSSR (Balkin, 2018b; Heldt, 2019b; Haupt, 2021). It was the first NSSR measure specifically targeting hate speech on social media, and has since 'set the international terms of engagement with online hate' (Tworek, 2021, p121). 25 countries have passed similar legislation, often explicitly modelled on NetzDG (Mchangama and Alkiviadou, 2020). NetzDG is therefore a 'prototypical case' of NSSR targeting online hate, as a representative example of a widely-used

approach (Hirschl, 2005). Additionally, as the earliest such measure, it offers the best opportunity for evaluation, since other initiatives' effects may not yet be apparent.

To assess NetzDG's effectiveness, I take an interdisciplinary sociolegal approach, drawing on empirical research to understand how the law functions in practice (Salter and Mason, 2007; Siems, 2009). This permits a normative evaluation of its success in achieving its primary policy aim (Taekema, 2018). NetzDG does not exclusively regulate hate speech, but aims to generally strengthen law enforcement online (He, 2020). However, the explanatory memorandum and government statements identify 'hate crime and other criminal content' as the problem targeted (Maas, 2017). German experts universally identify hate speech as the primary motivation (Echikson and Knodt, 2018; Heldt, 2019b; Wischmeyer, 2020).

Effectively addressing hate speech has two dimensions. First, *prevalence*: reducing the amount of hate content hosted, by deleting it or preventing users from posting, would reduce its effects. Second, these effects ultimately depend on *visibility*: how many people see a post is influenced by algorithmic recommendations and by users' social networks and interactions, in turn shaped by platform affordances. Accordingly, this paper seeks to evaluate how effectively NetzDG reduces both prevalence and visibility of hate speech. This is investigated through two linked research questions.

RQ1: How effectively does the regulatory model typified by NetzDG address factors known to drive the negative impacts of hate speech on social media?

This is investigated through an analysis of NetzDG's content, informed by the above empirical literature review. This helps contextualise NetzDG's regulatory strategy, highlighting choices to focus on certain solutions over others, and indicating some weaknesses.

However, social media platforms are complex sociotechnical systems; regulatory interventions may have unpredictable results (Helberger, 2020; Leerssen, 2020). Research should not only consider NetzDG's regulatory strategy in the abstract, but also examine available evidence about its effects to date. Some existing studies utilise platforms' transparency reports, but these give a very incomplete picture (see section 4(d)). Tworek and Leerssen (2019) suggest that inaccessible platform data makes any quantitative evaluation of NetzDG difficult. However, as section 2 shows,

qualitative research has greatly improved our understanding of online hate speech. This prompts my second research question.

RQ2: To what extent is NetzDG perceived as effective by CSOs campaigning against online hate speech?

I interviewed seven representatives of relevant CSOs, identified using a list of German anti-hate speech civil society initiatives (Das Nettz, n.d.). Interviews were semi-structured, using a standard questionnaire which used open questions to prompt reflections on NetzDG's objectives and effects (see Appendix II). Transcripts were coded inductively and compared using MAXQDA. Appendix I includes a pseudonymised list of interviewees with interview dates. Interview transcripts remain on file with the author.

This is a purposive sample of CSOs engaging with online hate speech, and is not representative of civil society perspectives. Additionally, the qualitative research design does not permit firm conclusions about how NetzDG affects the prevalence and visibility of hate speech. However, it offers insights into the informed perspectives of experts working with affected communities (who are themselves important actors in platform governance: Gorwa, 2019a). Since scholars identify community participation as a starting point for better platform governance (Duguay et al., 2020), this usefully complements previous research using transparency reports, which centre the perspectives of platform companies.

b. Content

In the lead-up to NetzDG's passage, online hate speech became a major political topic in Germany, due particularly to increasing online and offline racism following 2015's 'refugee crisis', as well as media discussions of hate speech and misinformation following the 2016 US elections, and concerns about similar dynamics in Germany's 2017 federal election (Gollatz and Jenner, 2018; Schulz, 2018). As comprehensively detailed by Gorwa (2021), NetzDG was spearheaded by then-justice minister Heiko Maas following the perceived failure of a self-regulatory code negotiated with Facebook, Twitter and Google. After being proposed in March 2017, it underwent significant amendment but was passed comparatively quickly before September's election, coming into full force from 1st January 2018 (Wischmeyer, 2020).

NetzDG applies to for-profit online platforms with over two million German users, which let users share content with others or the general public (§1(1)-(2)). Journalistic/editorial platforms, messaging services and platforms for 'specific content' (e.g. reviews) are excluded. Illegal content is defined according to 20 StGB provisions (see Appendix III). Any of these, if motivated by prejudice against a protected group, would qualify as hate speech under the definition adopted in section 2(a).

The key substantive provisions are §§2 and 3. §3 mandates new internal complaints-handling procedures. Platforms must offer an easily-recognisable, directly-accessible and permanently-available process for reporting illegal content. Staff handling complaints must have half-yearly training and support. Complaints-handling procedures must guarantee that:

- 'manifestly illegal' content is deleted or geoblocked in Germany (henceforth 'removed') within 24 hours
- all illegal content is removed expeditiously, generally within a week
- posters and complainants are immediately informed of decisions, with reasons
- removed content is saved for 10 weeks for potential legal investigations

Alternatively, platforms may refer removal decisions to approved self-regulatory institutions, although this option does not currently appear to be widely used (Eifert, 2020).

§2 creates new transparency obligations. All platforms receiving over 100 NetzDG complaints per year must publish a half-yearly report detailing:

- general efforts to prevent crime
- complaints-handling procedures
- number of complaints, broken down by source and grounds
- number of complaints leading to removal, within what timeframe
- information provided to posters and complainants
- organisation, staffing and training/support of complaints-handling departments
- membership of industry organisations

Under §4, breaching these obligations is a regulatory offence which can be fined up to €5 million by the *Bundesamt für Justiz* (BfJ), an agency within the justice ministry. Platforms based outside Germany must also designate a legal representative in Germany (§5).

NetzDG is not a criminal law, although it defines illegality by reference to the StGB. Nor does it alter platforms' intermediary liability for user-generated content. This is defined by the EU E-Commerce Directive (ECD) and German Telemedia Act (TMG). Under Article 14 ECD, platforms become liable for hosting unlawful content only once notified, unless they 'expeditiously' remove it. NetzDG aims to better enforce these existing obligations by mandating new internal procedures, backed by administrative penalties (Helberger, 2020). Liability arises only where platforms do not operate generally adequate reporting and complaints-handling systems: they cannot be fined for missing removal deadlines or wrongly deciding complaints.

c. Implementation and reform

As of 2021, eight platforms meet the two-million-user threshold and are subject to §§2 and 3: Change.org, Facebook, Instagram, Jodel, Reddit, SoundCloud, TikTok, Twitter and YouTube. Compliance has generally been high (BMJV, 2020). However, Facebook has been fined twice (Klausa, 2021): for using 'dark patterns' (Wagner et al., 2020) to mislead users into reporting content under its community standards instead of under NetzDG, and for excluding all such complaints from its transparency reports[6]. Several other enforcement procedures are underway against undisclosed companies for failing to appoint German representatives (Klausa, 2021).

In 2020, the government presented two reforms of NetzDG. The first, the Law Against Right-Wing Extremism and Hate Crime (GRH), aims to facilitate prosecutions regarding illegal content, complementing NetzDG's NSSR approach with stronger old-school speech regulation. As well as substantively updating criminal law (including explicitly criminalising sexual assault threats, a ubiquitous form of misogynist hate speech: Eckert, 2018), it added a new §3a NetzDG, requiring platforms to send removed illegal content to the BKA, with the poster's IP address. Due to privacy concerns, this was highly controversial. President Frank-Walter Steinmeier initially refused to sign the GRH due to doubts about its constitutionality. After some amendments, Steinmeier signed it

[6] Facebook's first report listed 886 NetzDG complaints, against over 200,000 each from Twitter and YouTube. Although Facebook was unusually obstructive in discouraging NetzDG complaints, most platforms preferentially review complaints under community standards, only then considering German law (often by a different team) if removal is not already indicated (Liesching, 2020; Eifert, 2020). This lets platforms apply common rules worldwide and use less resource-intensive decision-making procedures (Heldt, 2020; Wagner et al., 2020).

in March 2021. However, Google is currently challenging the law on constitutional rights grounds (Reuters, 2021); various CSOs and the Green Party also maintain that it is unconstitutional (Haufe, 2021; Bündnis 90/Die Grünen, 2021; Hiéramente, 2021).

The NetzDG Amendment Act (GÄNDG), promulgated in June 2021, makes four adjustments to NetzDG's regulatory model. First, transparency obligations are expanded: reports must detail the use of automated moderation; when complaints are decided under community standards versus NetzDG; and which groups are particularly likely to post or be affected by illegal content. Second, the BfJ's responsibilities are expanded to increase proactive compliance monitoring. Third, new procedural protections entitle posters and complainants to challenge decisions and submit disputes to BfJ-designated mediation services. Complainants must also be informed of the possibility of filing criminal complaints. Finally, to avoid conflict with the EU's updated Audiovisual Media Services Directive (AVMSD), video-sharing platforms based outside Germany are largely exempted from NetzDG.

d. Criticisms and evaluations

NetzDG was relatively popular: a representative 2018 survey found that 67% of Germans 'strongly approved' (Jacobs, 2018). However, it was highly controversial academically and politically. Objections mainly fall into three categories.

First, NetzDG is widely considered incompatible with the ECD (Liesching, 2020)[7]. It is argued to undermine harmonisation (Schulz, 2018), and to violate the country-of-origin principle, that platforms are regulated in the EU country where they are headquartered (Wischmeyer, 2020; Liesching, 2020)[8]. Second, critics raised procedural concerns about the speed of NetzDG's passage, its oversight by a non-independent ministerial department, and the possibility that it exceeds the federal government's competence (Heidrich and Scheuch, 2017; Schulz, 2018; Tworek and Leerssen, 2019). While important, these criticisms are quite particular to Germany. As this paper aims to draw generalisable conclusions about NetzDG's regulatory model, they are not examined in detail.

[7] Interviews by Gorwa (2021) indicate that this was also a common view within the European Commission, and at least one Commissioner wanted to formally oppose NetzDG. However, the Commission ultimately decided not to intervene, to avoid open conflict with Germany's government over a sensitive issue during an election year.
[8] Liesching notes that the GÄNDG exempted video-sharing platforms from NetzDG to respect the AVMSD's country-of-origin provisions, substantiating the argument that it contradicts the ECD's similar provisions.

More relevant is the third category: claims that NetzDG disproportionately restricts free speech, by delegating legal decisions to corporations, and by encouraging 'overblocking' of legal content (Heidrich and Scheuch, 2017; Echikson and Knodt, 2018; Tworek and Leerssen, 2019). This second point essentially reiterates Balkin's concerns about collateral censorship (see section 3(a)): where platforms are liable for users' speech, they will over-censor to minimise liability risks. On these bases, NetzDG faced international criticism (Kaye, 2017) and sustained opposition from German journalists, lawyers, startups and civil society (Tworek and Leerssen, 2019).

In my view, concerns about privatising legal decisions are overstated. The scale of online platforms makes it implausible that first-instance moderation decisions could be made by state institutions (Eifert, 2018; Wischmeyer, 2020; douek, 2021). Such detailed state supervision of online communications would also raise major free speech concerns. Platforms already exercise extensive control over users' speech for commercial purposes, often arbitrarily censoring minorities and political speech (Klonick, 2018; Cobbe, 2020; York, 2021). In this context, procedural regulations governing illegal hate speech are hardly the biggest threat to free speech (Heldt, 2019a; Wischmeyer, 2020). Indeed, the now-amended NetzDG gives posters stronger procedural safeguards than standard contractual relationships with platforms (Liesching, 2020), which typically offer no recourse against censorship (Leerssen, 2015).

How far overblocking concerns have materialised is debated. Overblocking in copyright cases has been documented extensively (Keller, 2021). Whether this applies under NetzDG remains unclear. Eifert (2018) argues collateral censorship is inapplicable, since (unlike in copyright law) platforms are only liable for systematic procedural failures, not individual illegal posts. Indeed, since §3(2)(1) requires platforms to 'check whether the content reported in the complaint is unlawful'[9], systematically removing content without genuinely considering legality would arguably violate this obligation just as much as systematic under-removal. It could also violate the developing German jurisprudence on horizontal effect of constitutional rights against platforms, which has established that contractual terms authorising content moderation are to be interpreted – and potentially invalidated – in accordance with constitutional rights. Germany's highest civil court recently held that Facebook's terms and conditions are invalid to the extent that they strike an unfair balance between the parties' rights and interests, including users' rights to free expression (*Bundesgerichtshof Urteil des III. Zivilsenats vom 29.7.2021*). A lower court has held that Facebook's terms do not permit arbitrary or unreasonable content moderation (*Oberlandesgericht Braunschweig Urteil vom 05.02.2021*).

[9] Official translation (BMJV, 2017).

Nonetheless, liability for systematic under-removal could plausibly create a bias towards removal in doubtful cases. Transparency reports are not detailed enough to draw firm conclusions. They indicate, broadly, that complaints are typically resolved within 24 hours and around a quarter result in removal; this has been treated as evidence both for (Liesching, 2020) and against overblocking (Eifert, 2020). A Counter Extremism Project investigation (2020) found that platforms (especially YouTube) still frequently fail to remove manifestly illegal reported content. However, there is anecdotal evidence of spurious NetzDG complaints resulting in blocking (Delcker, 2020; Shephard, 2020). Given biases in human and automated moderation (see section 3(b)), it is possible that overblocking does not happen at scale but does affect certain communities or content types.

Compared to free speech concerns, surprisingly little research examines NetzDG's effectiveness as a policy measure. Researchers have focused primarily on the effectiveness of the transparency obligations in strengthening accountability – generally regarded as low (Heldt, 2019b; Tworek and Leerssen, 2019; Wagner et al., 2020). Yet NetzDG's explanatory memorandum and government statements indicate that promoting transparency was secondary to the overriding aim of preventing hate speech; how effectively it achieves this deserves more attention. Plausible concerns about overblocking only make this more important. German law requires limitations of constitutional rights to be proportionate – meaning they pursue a legitimate aim, are suitable to achieve it, are the least intrusive measure available, and appropriately balance competing interests (Degenhart, 2015). NetzDG's effectiveness against hate speech is highly relevant in assessing suitability and appropriate balancing.

A detailed analysis by Eifert (2020), commissioned by the BMJV, concluded that NetzDG's objectives were 'in large part achieved'. Eifert analysed transparency reports and surveyed platforms, the BfJ, legal experts and civil society (primarily legal professional associations). While he gives an authoritative account of NetzDG's implementation, it does not support the aforementioned conclusion. Eifert notes that any effects on the prevalence and visibility of online hate speech remain uncertain, but concludes, 'The law significantly improved the complaints-handling and public accountability of network providers in handling specified illegal content. Network providers largely implemented NetzDG's key requirements…In this way, the aims pursued were in large part achieved' (p151, own translation). Equating compliance with policy objectives is misleading. Improving complaints-handling procedures was not an end in itself, but a means of reducing hate speech. Eifert's evidence does not establish that this was achieved.

Heldt (2019b, 2020), Echikson and Knodt (2018) and Tworek and Leerssen (2019) offer independent evaluations based on transparency reports. They concur that compliance is generally high, but also consider whether NetzDG effectively incentivises faster removal of hate content. The most detailed assessment is from Heldt (2019b). She concludes that reporting obligations have not extracted much useful information, but some indications suggest more hate speech is being removed. All platforms examined devoted more resources (including legal experts) to examining complaints, and handled most within 24 hours. Hate speech was the most common complaint category, suggesting these new procedures improved how it is moderated.

However, without more granular and comprehensive information, firm conclusions cannot be drawn. Transparency reports only include content reported by users, which is neither exhaustive nor representative of online hate speech. As Tworek and Leerssen (2019, p7) note, 'it will require much more research — and greater access to data — to determine whether NetzDG is achieving its aim, and whether any benefits outweigh the harms to free speech'. The following sections aim to begin answering this question.

5. A critique of the NetzDG model

a. Scope

NetzDG's scope is limited in two important ways. First, the two-million-user threshold may limit its impact, as people seeking or sharing hate speech can move to smaller platforms. Three interviewees considered this to seriously limit NetzDG's effectiveness:

> 'Digital violence is also happening there…it isn't always the size of the platform.' (Interviewee A)

> 'It's not considering smaller platforms where we know that hate is organised…What we see on the bigger networks like YouTube and Facebook and Twitter is just the execution of plans made on smaller platforms.' (Interviewee F)

However, since compliance with NetzDG requires non-trivial investments of resources and personnel, this could be justified by avoiding heavy costs which would disadvantage smaller

companies in an already uncompetitive market (Gillespie and Aufderheide, 2020). Whether NetzDG strikes the right balance between these considerations is open to debate; arguably some kind of tiered structure with less stringent obligations for smaller platforms (as in the EU Digital Services Act (DSA)) would be more appropriate than the sharp two-million-user cut-off.

§1(1) also excludes certain platform types. Editorial/journalistic platforms are already regulated through media law, and host less user-generated content, which makes their exclusion seem reasonable. Excluding non-profit platforms can also be justified by compliance costs, while excluding messaging services seems justified on privacy grounds, and by the limited social impact of hate speech in private communications[10]. However, the GÄNDG's exclusion of video-sharing platforms (see section 4(c)) is concerning. YouTube is an important channel for far-right networks and (unlike other big platforms) enables direct monetisation of extremist content, incentivising creators to continue producing it (Lewis, 2020; Munger and Phillips, 2020).

Another important limitation is that all obligations in §§2 and 3 relate only to content violating one of the 20 listed StGB provisions (see Appendix III). Some interviewees suggested that using StGB provisions to define what is banned has enabled well-informed users to avoid moderation by staying just on the legal side. This illustrates a practical disadvantage of relying on strictly-defined categories to govern complex sociotechnical systems where user behaviour can change in response to regulation. Moreover, as section 2(b)(i) showed, legal content can encourage criminal hate speech, circulate it, and spread hateful ideologies. This calls for a more holistic approach, which also considers appropriate ways to reduce the impacts of harmful or discriminatory content which is not illegal.

Regulating how platforms treat legal content may seem intrusive into media independence and free expression. However, as section 3(c) notes, this could include interventions like downranking and content-neutral design changes which are less restrictive than censorship[11]. All social media content is subject to intervention, as platforms actively construct users' information environments for commercial purposes (Carmi, 2020; Pasquale, 2020). In this context, regulations incentivising interventions to address hate speech are not intrinsically objectionable. Indeed, leaving platforms'

[10] Concerns were raised about the earlier interpretation of this provision to include Telegram, a popular messaging service which includes public 'channels' and is widely used by Germany's far right (Fielitz et al., 2020, p64); this was also mentioned by interviewees. However, the BfJ now considers Telegram covered by NetzDG and is currently pursuing regulatory action against it for failing to name a German legal representative (Stenner and Reuter, 2021).
[11] douek (2021) suggests content governance using a wider range of interventions will be more protective of free speech, as censorship can then only be used where absolutely necessary.

private opinion power unregulated is also unlikely to serve media freedom or free speech (Helberger, 2020). The question is how regulation can provide accountability without excessive state interference, and reduce the impacts of harmful speech without completely censoring it.

b. Substantive obligations

i. Overview

In this respect, a major weakness of NetzDG is immediately apparent. §2's transparency reporting obligations and §3's expanded moderation procedures implement a narrowly censorship-focused version of NSSR, imposing a binary approach to moderation whereby content is either illegal, and must be removed, or legal, in which case no action is required. They fail to address broader considerations around how platforms' sociotechnical environments and design choices encourage and disseminate hate speech.

Generally, interviewees were pessimistic about NetzDG's overall regulatory orientation. All but one interviewee saw no evident reduction in the prevalence or visibility of hate speech. Three were very critical, calling NetzDG 'not a good idea' (Interviewee C) or 'a really crappy legislation' (Interviewee A). In contrast, four were qualifiedly positive, describing it as a positive step which shows legislators take hate speech seriously. However, even interviewees with generally positive views described NetzDG as mostly symbolic, or as driven more by political pressure and PR concerns than concrete policy goals. There was a general view that platforms may 'follow the letter of the law' (Interviewee G), but that NetzDG has failed to achieve the systemic changes to platform governance which would effectively reduce hate speech.

ii. Transparency

Transparency obligations appear to support the effective implementation of NetzDG in the narrow sense of enforcing platform compliance. Transparency is often considered a weak form of regulation which demands little concrete change (Ananny and Crawford, 2018; Gorwa and Garton Ash, 2020). However, by combining transparency reports with substantive obligations and significant sanctions, NetzDG creates a useful enforcement tool. This is borne out by the BfJ's enforcement action against Facebook, which was based on inadequate reporting, but also demanded substantive changes in complaints-handling practices. Wagner et al. (2020) suggest this

could discourage similar dark patterns in future, strengthening the effectiveness of NetzDG's complaints-handling procedures.

However, it is clear from previous research and from my interviews that NetzDG reports are inadequate to assess the prevalence and visibility of hate speech, or NetzDG's effects thereon: they lack detail, standardisation across companies, and independent oversight. Interviewee C further noted that their relevance is limited by the reliance on user reporting: they give no indications of how many people encountered hate speech but did not report it. More comprehensive transparency obligations would not only aid independent research, but would also incentivise more effective measures by platforms, by exposing them to evidence-based public criticism for failing to address hate speech[12]. This could include greater transparency regarding recommendations and other governance measures.

iii. Complaints-handling

Platforms' new obligations under §3 NetzDG, to quickly examine user complaints about illegal content and delete it if appropriate, may have some positive impact. As section 3(b) shows, moderating hate speech reduces its audience and the incentives to produce it (Munger and Phillips, 2020), affecting both prevalence and visibility. Germany's intelligence service claims to have observed exactly these effects in Germany's far-right since NetzDG's introduction (NTV, 2019), although Tworek and Leerssen (2019) question the evidence for this claim. Although six of my interviewees thought NetzDG has not reduced the prevalence or visibility of hate speech at all, one argued that it has achieved faster removals:

> 'I work with a lot of international partners, and if you see the situation in other European countries, for example, I see a difference. So at least very explicit death threats or other crimes are taken down really fast in Germany, and really slow in other countries.' (Interviewee D)

While she did not think the overall prevalence of hate content had decreased, Interviewee F agreed that takedowns were faster, which she said victims appreciated. Interviewees also highlighted some indirect positive effects, suggesting that by starting a public debate about hate speech and showing

[12] The GÄNDG aims to address this by requiring more detailed reports, but they will still only cover content reported by users, and lack independent oversight.

political will to address it, NetzDG pressured platforms into investing more resources in moderation and hate speech prevention. These effects should not be disregarded, although it is difficult to disentangle NetzDG's effects from other social and political factors. Platforms certainly increased their investment in moderation and expert personnel in Germany after NetzDG's introduction (Oltermann, 2018; Heldt, 2019b).

Nonetheless, §3 has two major flaws. First, as indicated above, focusing exclusively on moderation significantly limits its effectiveness. Identifying and moderating all hateful content after it is posted is not feasible. On the other hand, factors like algorithmic recommendations, platform architecture and social/interactive affordances all influence *in advance* the likelihood that users will post or view hate content. Intervening at these earlier stages is therefore crucial.

However, NetzDG does not address these factors at all. To take two examples mentioned in section 2(b)(ii), it does nothing to prevent Facebook from actively recommending far-right groups, or algorithmically promoting 'hate bait' posts which encourage hate speech while staying just within the law. Several interviewees noted the importance of design features – especially recommendations – in spreading hate speech, and criticised their omission. Some suggested that effectively addressing hate speech would require changes to business models and the optimisation of design and algorithms for profit:

> 'They are built on a system to exclude, to bring people to rage…it's a money machine, built on racism, sexism.' (Interviewee A)

Given the limitations of moderation as a solution, the absence of incentives to pursue other interventions means NetzDG's effectiveness will be limited. If anything, by creating large financial incentives to focus on censoring illegal content, it may actively disincentivise platforms from expending additional resources on alternative solutions.

In particular, one interviewee argued that NetzDG's obligations just encourage platforms to hire more lawyers and 'policy guys' to manage illegal content, while software engineers (a scarcer, more expensive resource) can still be directed exclusively to revenue-generating activities. He suggested design solutions would more effectively restrict hate speech, and that NetzDG's attempt to consider content governance entirely separately from platform design and business models is unlikely to succeed:

> 'The business part [of platforms] where the expertise is, where these thousand engineers are, is separated from this compliance and policy part, that seems to be more operating like a PR department, managing folks like myself. I think that is because it is less costly to hire ten policy guys…much less costly than hiring ten engineers to actually fix the problem.' (Interviewee G)

The narrow focus on censorship is also closely connected with the scope of the law and its exclusive focus on illegal content. If censorship is the only intervention considered, free speech concerns mean it can only be mandated in strictly limited circumstances, and its impacts will necessarily be limited. A more holistic approach to platform governance, utilising other measures which could discourage hateful content without censoring it completely, could engage more effectively with the full range of factors driving online hate speech.

Moreover, §3's moderation requirements relate exclusively to content reported by users. This further limits their practical relevance, since much hate speech goes unreported. Contrary to some assumptions, NetzDG does not mandate or encourage automated moderation (Heldt, 2019b); nor does it encourage any manual proactive detection of illegal content. Platforms' incentives are to direct maximum resources towards handling complaints efficiently, not to search unreported content. This is substantiated by the Counter Extremism Project's (2020) investigation, which found that Facebook ignored manifestly illegal pictures in the same album as those reported, while YouTube ignored identical videos. The absence of legal or financial incentives for proactive measures can explain this. Reporting is also easily abused to silence victims of discrimination. NetzDG complaints have frequently been used in this way (Delcker, 2020; Shephard, 2020). This is also not surprising: platforms' incentives are to process reports quickly, rather than investigating other forms of discrimination and harassment, like coordinated malicious reporting.

In addition to these practical issues, relying on user reporting raises more fundamental concerns. As Duguay et al. (2020) show, it places the onus of dealing with hate speech on victims, which may be demanding, for example in mass coordinated harassment situations. It also promotes an individualistic understanding of hate speech, and social media use generally: users are responsible for curating their own online experiences, while platforms disclaim responsibilities to create a safe environment (Siapera and Viejo-Otero, 2021). In interviews, the burden of user reporting emerged as a central point of criticism:

> 'You have only the possibility at the moment, still, to work on this yourself, if you can…which is really difficult for the victims themselves.' (Interviewee D)

As well as criticising this fundamentally individualistic approach, interviewees highlighted specific problems with how reporting has been implemented. They criticised NetzDG for failing to prevent platforms from creating unnecessarily complicated interfaces, and for expecting users to be familiar with technical legal provisions (also criticised by Eifert, 2020):

> 'It's always long, really a lot of clicks you have to do. So it's not a really nice user experience, it's more like, "Don't do it."' (Interviewee A)

> 'It's totally threatening, that whole way it's done. I know it should be three clicks…no more. One, two, send. That's what we know for how digital things are effective.' (Interviewee C)

> 'It's a little bit frightening to see this form, and they're asking you, "Are you really sure that this is chargeable with a criminal offence?" And I'm not an attorney!' (Interviewee A)

As Crawford and Gillespie (2016) observe, flagging is a 'thin' form of participation which offers users little meaningful participation in content governance. Multiple interviewees commented that platforms remain highly intransparent, and that victims of hate speech have few avenues for meaningful communication with platforms. NetzDG allows users to express opinions on whether individual posts are criminal, but not on general moderation policies, other possible interventions, or broader changes in platform governance. Requiring platforms to invest more staff and resources in their existing content moderation systems without otherwise reforming these systems entrenches a top-down governance model where users' views and experiences are marginal.

This is not just a problem for individual victims, but undermines the search for policy approaches which serve communities affected by hate speech. Interviewees generally considered that NetzDG's reporting-based model was not informed by these communities' experiences or those of civil society experts:

> 'The situation for victims of digital violence didn't change at all…they didn't get anything to solve their situation better.' (Interviewee D)

> '[Groups] most affected from digital violence – this NetzDG does nothing for them.' (Interviewee A)

Overall, NetzDG makes only incremental adjustments to the model of content moderation – content flagged by users is queued for review under standardised rules – that already predominated within large commercial platforms (Klonick, 2018). As Wischmeyer (2020, p42) notes, the §3 obligations 'are mostly common-sensical and build on the technological infrastructures which most intermediaries have already set up'. Seemingly, these pre-existing systems have substantially constrained the legislator's imagination. NetzDG misses the opportunity to require more radical changes to how platforms govern content and user interactions, which would align these processes with the needs of people affected by hate speech.

c. Social and institutional context

In addition to these flaws in NetzDG's regulatory strategy, interviewees highlighted aspects of the broader policy context which undermine its effectiveness. The need for better support for victims was a recurring theme. Three interviewees said police are often unhelpful:

> 'We need more training for the police force, because they also don't know what digital violence is. They always say, "Turn it off and go home, there's nothing here."' (Interviewee A)

Three discussed NGOs supporting victims of hate speech, which are overstretched and underfunded. Increasing their funding was described as another overlooked but potentially effective intervention.

Another aspect was public awareness of NetzDG, and of laws around hate speech more broadly. Interviewee C suggested that many people are unaware of reporting procedures, limiting their practical relevance. Other interviewees suggested that those posting hate speech are unaware of or indifferent to potential consequences. They linked this to the lack of concrete consequences other than content deletion, and the unlikelihood of prosecutions. Despite the name 'Network Enforcement Act', interviewees generally considered that NetzDG has done little to actually strengthen criminal law enforcement online:

> 'If something is illegal, we delete it, and then that's it. Nothing else happens, there are no consequences for the person that posts illegal content.' (Interviewee F)

> 'Many people still don't believe that crimes they commit online can also lead to charges.' (Interviewee E)

Three interviewees suggested NetzDG's moderation-focused approach may actually have made this problem worse, as rapidly removing content can make it more difficult to pursue criminal complaints. Another problem identified by two interviewees was inadequate funding and training for police and prosecutors to investigate online hate speech and offer support to victims – again, something NetzDG does not address at all[13].

There was also a consensus around the need for a broader spectrum of policy interventions to tackle the overall social context that produces hate speech, rather than focusing exclusively on platform governance and speech regulation:

> 'Rules and regulation are one thing, but if you don't change the mindset of the community or of the society, you can have a thousand rules, it won't change anything.' (Interviewee A)

> 'I think there's a lot that society can do. Anything that's related to what we do or what the NetzDG does is – well, criminalises things that have already been said. We are there when things are already too late.' (Interviewee B)

This echoes the arguments made above about the limitations of moderation, and the ineffectiveness of focusing only on deleting hate content, rather than proactive prevention. The interviews suggest that a preventive approach should take a broader perspective, considering not only all available platform governance mechanisms, but also relevant policy interventions in education, criminal justice, and victim support.

6. Recommendations and future research

[13] The GRH aims to improve enforcement by providing user data to the BKA, but has raised fundamental rights concerns: see section 4(c). It does not extend support for victims.

a. Systemic and preventive regulation

This examination of NetzDG's content and reception by CSOs exposes the limitations of this regulatory model. As discussed in section 4(a), government statements indicate that the strategy behind NetzDG was to tackle hate speech by strengthening the enforcement of specified criminal law provisions (through mandating deletion of criminal content), rather than to address broader questions about how social media environments can facilitate harmful behaviour. Such a narrow strategy, which only requires intervention after criminal speech takes place and ignores all contextual factors leading up to it, is intrinsically limited.

That is especially the case in the social media context. There is a fundamental mismatch between NSSR mandating censorship of narrowly-defined content categories, and the complex sociotechnical environments of large-scale, networked, algorithmically-curated social media, which influence user interactions in myriad ways. Moreover, in democracies, censorship-based NSSR can only legitimately be used in very limited circumstances. Yet legal content can encourage hate speech – and all such content is already subject to extensive intervention at platforms' discretion, through recommendations and design choices which enable, encourage or constrain certain interactions. There is no reason these private governance mechanisms should be totally unregulated; but censorship-based NSSR is not the appropriate regulatory strategy. Effectively addressing hate speech demands more creative regulatory solutions, which would incentivise companies to design their platforms and recommendation algorithms to discourage and reduce the impacts of hate speech.

NetzDG incrementally adjusts failing moderation systems, rather than demanding new approaches – even though there is extensive evidence on the nature of the problem, and many constructive proposals for effective responses. That NetzDG ignores all aspects of platform governance other than moderation – as well as broader social policy questions like education – suggests it was not based on a thorough understanding of the problem. This can partly be explained by its political context. As Gorwa (2021) shows, time constraints related to procedural requirements and Germany's electoral timetable favoured the quick passage of a law regarded by many policymakers as flawed. This likely favoured simple, incremental changes over further-reaching reforms. Another factor could be the predominance of legal professionals in Germany's civil service, especially the BMJV (Wegrich and Hammerschmid, 2018). NetzDG shows continuities with German legal traditions (He, 2020; Tworek, 2021), and reflects a legalistic understanding of content governance,

based on clear rules and statutory categories. Its drafters' professional background may have favoured this approach over engineering-based solutions.

Although my analysis suggests several small reforms which could marginally increase NetzDG's effectiveness (e.g. mandating more user-friendly reporting systems), such an incrementalist approach would be misguided. NetzDG's regulatory model is fundamentally flawed. Effective regulation must be more *systemic*, considering the whole platform environment and how users interact with it instead of banning individual posts, and more *preventive*, proactively discouraging people from posting or viewing hate speech rather than only deleting it afterwards.

In this regard, douek's (2021) nuanced account of NSSR – which retains Balkin's triadic model, but argues for systemic content governance strategies including many interventions other than censorship – seems more promising. However, legislators should not only regulate 'remedies' for particularly harmful content (douek's primary focus), but should also address the methods and objectives of content governance more generally. Through recommendations and technical affordances, platforms' opinion power shapes users' experiences, information environments, and – ultimately – behaviour. Incentivising the use of such power to actively discourage and suppress hate speech, instead of promoting it, should be the goal of regulation.

Given the complexity and diversity of platform architectures and user cultures, understanding how to achieve this is a challenge for regulators and researchers. However, while interventions in complex sociotechnical systems may not always have predictable results, they can be steered in desired directions. Currently, platforms are effectively optimised for ad impressions and profit (Cobbe, 2020). Regulation should ensure design and optimisation processes also consider public interest objectives like discouraging hate speech.

One approach would be procedural regulations requiring internalisation of public interest considerations in design and governance processes, such as mandatory risk assessments and audits. Such regulations have practical advantages over prescriptive rules (e.g. accommodating diversity between platforms) and alleviate free speech concerns raised by direct state intervention in content governance (douek, 2021). Platforms sometimes test such design changes, but these initiatives are ad hoc, intransparent and easily overridden by profit imperatives (recently illustrated by two separate investigations showing that Facebook executives overruled product changes aiming to

discourage hateful and divisive content on profit grounds: Hao, 2021 and Hagey and Horwitz, 2021). They should be systematic, transparent and obligatory, even where they affect profits.

Legislators could also more radically reform the structures of platform governance. As my interviews highlighted, profit- and engagement-based business models conflict with policies which would require platforms to de-amplify engaging content, divert expensive engineering resources from revenue-generation, and prioritise users' interests over shareholders'. Noble (2018) argues that platforms which organise information based on profitability will never serve marginalised groups. Conversely, participation by users and marginalised communities is identified as an important starting point for effective anti-hate speech measures (Duguay et al., 2020). My analysis shows that the NetzDG model does not respond to the needs and perspectives of victims or civil society; giving these groups more input would likely have produced a different strategy. Promoting participatory, democratic and non-commercial platform governance structures could open new possibilities to prevent hate speech and create a more equal, inclusive online public sphere.

Finally, social media are a distinctive information environment, and many changes could and should be made in platform regulation to tackle hate speech. However, interviewees emphasised that online hate speech should not be considered in isolation from broader social factors. On-platform solutions are important, but cannot replace law enforcement, social work and education programmes which address underlying discrimination and prejudice. The regulatory strategies discussed here would be most effective as part of a broader anti-discrimination programme encompassing these other policy fields.

b. Future research

This analysis has highlighted the flaws of the NetzDG model and suggested alternative strategies. However, many questions remain. What concrete design changes can different platforms make? What regulatory and governance structures could incentivise such changes? How can the perspectives of users, marginalised groups, and CSOs inform these strategies? These questions require further research in disciplines including human-computer interaction and communications studies as well as law, regulation and governance. Platforms and independent researchers should be systematically and transparently investigating potential interventions; external access to data is vital (Matias, 2016).

Future legal research could look to other fields to inform regulatory strategies. For example, Helberger (2020) suggests media law – which has long tried to steer a course between unaccountable private power and excessive state intervention – offers useful inspiration. Regulatory studies literature could illuminate how law influences platforms' governance processes, and how procedural regulations could provide effective accountability (Gorwa, 2020).

In addition, more research is needed on the international implications of, and constraints on, such regulations. Moderation rules can easily be implemented in individual countries through geoblocking, but systemic design and governance changes have global ramifications – recently illustrated by changes several platforms introduced globally in response to the UK's Age-Appropriate Design Code (Hern, 2021). Such regulations might face opposition from the US, home to most major platforms. Like NetzDG, they might inspire similar regulation by authoritarian countries, with severe consequences for free speech and media freedom. EU member state regulations might also risk incompatibility with the country-of-origin principle; harmonised EU-level reform would be more legally straightforward, though politically complex.

Finally, the DSA proposals incorporate various procedural obligations along the lines discussed here, but leave untouched the overall platform governance structure that Leerssen (2020, p50) characterises as 'regulated oligopoly'. Further research is needed to investigate its likely impact on hate speech, discrimination and equality.

7. Conclusion

This paper aimed to evaluate how effectively the 'NetzDG model' of NSSR addresses online hate speech, through an empirically-informed analysis of the legal text and its regulatory strategy. It shows that NSSR which focuses on censoring narrowly-defined content categories fails to address the factors driving hate speech on social media, which include not only underlying social prejudices but also platforms' technical features and the behaviours and interactions they encourage. The findings support calls by scholars like Helberger, Cobbe and douek for more systemic approaches to social media governance, which challenge the power of commercial platforms.

Effective anti-hate speech regulations must address content governance holistically, rather than targeting individual posts, and ensure platforms' opinion power is exercised in the public interest. In short, we need a *new* new school of speech regulation: one which retains the triadic model

whereby states regulate how private intermediaries exercise their power over public discourse, but takes a broader view of what this power comprises. Power does not simply come from deciding what users can say: by controlling the availability of information and constructing the sociotechnical contexts of their interactions, platforms influence what people want to say or hear in the first place. Effective hate speech regulation must actively ensure that this power is exercised responsibly. Though NetzDG has been influential internationally, it is to be hoped that future regulations like the DSA will recognise its limitations and pursue a more radical regulatory strategy.

Bibliography

Abé, N., Hoffmann, H. and Peteranderl, S. (2021, February 14). „Es ist ein organisierter Lynchmob." *Der Spiegel*. Retrieved September 20 2021, from https://www.spiegel.de/ausland/digitale-gewalt-gegen-politikerinnen-es-ist-ein-organisierter-online-lynchmob-a-48c5a67c-a20f-4a59-8d54-d52415fb1a4e

Ananny, M. (2019). *Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance*. Knight First Amendment Institute. Retrieved May 10 2021, from https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance

Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973-89. https://doi.org/10.1177/1461444816676645

Balkin, J.M. (2014). Old-School/New-School Speech Regulation. *Harvard Law Review, 127*(2296), 2329-41.

Balkin, J.M. (2016). Information Fiduciaries and the First Amendment. *UC Davis Law Review, 49*(4), 1183-234.

Balkin, J.M. (2018a). Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *UC Davis Law Review, 51*, 1149-210.

Balkin, J.M. (2018b). Free Speech is a Triangle. *Columbia Law Review*, *118*(7), 2011-2056.

Balkin, J.M. (2018c). *Fixing Social Media's Grand Bargain* (Aegis Series Paper No. 1814). Hoover Institution. Retrieved May 10 2021, from https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf

Balkin, J.M. (2020). *How to Regulate (and Not Regulate) Social Media.* Knight First Amendment Institute. Retrieved May 10 2021, from https://knightcolumbia.org/content/how-to-regulate-and-not-regulate-social-media

Barak, A. (2005). Sexual harassment on the Internet. *Social Science Computer Review*, *23*(1), 77-92. https://doi.org/10.1177/0894439304271540

Barnidge, M., Kim, B., Sherrill, L.A., Luknar, Ž. and Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and Informatics*, *44*. https://doi.org/10.1016/j.tele.2019.101263

Ben-David, A. and Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication, 10*, 1167–1193. https://ijoc.org/index.php/ijoc/article/view/3697

Berger, J.M. and Perez, H. (2016). *The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters.* George Washington University Program on Extremism. Retrieved May 10 2021, from https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/JMB%20Diminishing%20Returns.pdf

Bowers, J. and Zittrain, J. (2020). Answering Impossible Questions: Content Governance in an Age of Disinformation. *Harvard Kennedy School Misinformation Review, 1*(1). https://doi.org/10.37016/mr-2020-005

Bucher, B. (2020, October 30). WhatsApp, WeChat and Facebook Messenger Apps – Global useage of Messaging Apps, Penetration and Statistics. *Messenger People.* Retrieved May 10 2021, from https://www.messengerpeople.com/global-messenger-usage-statistics/#:~:text=According%20to%20a%20study%20by,Germany%20(Statista%20April%202019)

Bundesgerichtshof, *Urteile des III. Zivilsenats vom 29.7.2021* (III ZR 179/20 und III ZR 192/20), https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2021/2021149.html

Bundeskriminalamt (2020a). *Politisch motivierte Kriminalität im Jahr 2019*. Bundesministerium des Innen, für Bau und Heimat. Retrieved May 10 2021, from https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2020/pmk-2019.pdf?__blob=publicationFile&v=11

Bundeskriminalamt (2020b). *Politisch motivierte Kriminalität (PMK) – rechts.* Bundesministerium des Innen, für Bau und Heimat. Retrieved May 10 2021, from https://www.bka.de/DE/UnsereAufgaben/Deliktsbereiche/PMK/PMKrechts/PMKrechts_node.html

Bundesministerium für Justiz und Verbraucherschutz (2015). *Gemeinsam gegen Hassbotschaften: Ergebnispapier der Task Force ‚Umgang mit rechtswidrigen Hassbotschaften im Internet' vorgeschlagene Wege zur Bekämpfung von Hassinhalten im Netz.* Retrieved May 10 2021, from https://www.bmjv.de/SharedDocs/Downloads/DE/News/Artikel/12152015_TaskForceErgebnispapier.html

Bundesministerium für Justiz und Verbraucherschutz (2020). *Bericht der Bundesregierung zur Evaluierung des Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG).* Retrieved May 10 2021, from https://www.bmjv.de/SharedDocs/Artikel/DE/2020/090920_Evaluierungsbericht_NetzDG.html

Bundesministerium für Justiz und Verbraucherschutz (2021). *Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act).* Retrieved May 17 2021, from https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html

Bundesverfassungsgericht, *Beschluss des Ersten Senats vom 27.5.20* (1 BvR 1873/13), http://www.bverfg.de/e/rs20200527_1bvr187313.html

Bündnis 90/Die Grünen (2021, April 15). *Gesetz Gegen Hasskriminalität.* Grüne Bundestag. Retrieved May 10 2021, from https://www.gruene-bundestag.de/themen/rechtspolitik/gesetz-gegen-hasskriminalitaet-umgehend-verfassungskonform-machen

Butler, A. and Parrella, A. (2021, May 5). *Tweeting with consideration.* Twitter Blog. Retrieved May 10 2021, from https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration.html

Carmi, E. (2020, December 14). The Organic Myth. *Real Life*. Retrieved May 10 2021, from https://reallifemag.com/the-organic-myth/

Citron, D.K. (2014). *Hate Crimes in Cyberspace*. Cambridge, Massachusetts: Harvard University Press.

Cobbe, J. (2020). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00429-0

Cobbe, J. and Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles. *European Journal of Law and Technology, 10*(3). https://ejlt.org/index.php/ejlt/article/view/686

Connolly, K. (2020, October 6). Hundreds of rightwing extremist incidents by German security services revealed. *Guardian*. Retrieved May 10 2021, from https://www.theguardian.com/world/2020/oct/06/report-reveals-hundreds-of-rightwing-extremist-incidents-by-german-security-services

Connolly, K. (2021, May 4). German society 'brutalised' as far-right crimes hit record levels. *Guardian*. Retrieved May 10 2021, from https://www.theguardian.com/world/2021/may/04/rightwing-extremism-germany-stability-interior-minister-says

Copland, S. (2020). Reddit quarantined: can changing platform affordances reduce hateful material online? *Internet Policy Review, 9*(4). https://doi.org/10.14763/2020.4.1516

Counter Extremism Project (2020). *NetzDG 2.0: Recommendations for the amendment of the German Network Enforcement Act (NetzDG) and Investigation into the actual blocking and removal processes of YouTube, Facebook and Instagram*. Counter Extremism Project. Retrieved May 10 2021, from https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper%20April%202020%20ENG.pdf

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

Daniels, J. (2018). The algorithmic rise of the 'alt-right'. *Contexts*, *17*(1), 60-65. https://doi.org/10.1177/1536504218766547

Das Nettz (n.d.). *Initiativen gegen Hass im Netz – Wer engagiert sich wie?* Das Nettz. Retrieved December 28 2020, from https://www.das-nettz.de/initiativen-gegen-hass-im-netz-wer-engagiert-sich-wie

Davis, T., Livingston, S. and Hindman, M. (2019). *Suspicious Election Campaign Activity on Facebook: How a Large Network of Suspicious Accounts Promoted Alternative für Deutschland in the 2019 EU Parliamentary Elections*. Institute for Data, Democracy and Politics. Retrieved May 10 2021, from https://smpa.gwu.edu/sites/g/files/zaxdzs2046/f/2019-07-22%20-%20Suspicious%20Election%20Campaign%20Activity%20White%20Paper%20-%20Print%20Version%20-%20IDDP.pdf

Degenhart, C. (2015). *Staatsrecht I: Staatsorganisationsrecht* (31st ed.). Heidelberg: C.F. Müller.

Delcker, J. (2020, February 24). Germany's balancing act: Fighting online hate while protecting free speech. *Politico*. Retrieved May 10 2021, from https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/

douek, e. (2021). Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review, 121*(3), 759-833. https://www.columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/

Duarte, N., Llanso, E. and Loup, A. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy and Technology. Retrieved September 24 2021, from https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

Duguay, S., Burgess, J. and Suzor, N. (2020). Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence*, *26*(2), 237-252. https://doi.org/10.1177/1354856518781530

Echikson, W. and Knodt, O. (2018). *Germany's NetzDG: A key test for combatting online hate* (CEPS Research Report No. 2018/09). CEPS. Retrieved May 10 2021, from http://aei.pitt.edu/95110/1/RR_No2018-09_Germany's_NetzDG.pdf

Eckert, S. (2018). Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. *new media & society, 20*(4), 1282-1302. https://doi.org/10.1177/1461444816688457

Eifert, M. (2018). Das Netzwerkdurchsetzungsgesetz und Plattformregulierung. In Eifert, M. and Gostomzyk, T. (Eds.), *Netzwerkrecht: Die Zukunft des NetzDG und seine Folgen für die Netzwerkkommunikation* (1st ed., pp.9-44). Baden-Baden: Nomos.

Eifert, M. (2020). *Evaluation des NetzDG im Auftrag des BMJV*. Bundesministerium für Justiz und Verbraucherschutz. Retrieved May 10 2021, from https://www.bmjv.de/SharedDocs/Downloads/DE/News/PM/090920_Juristisches_Gutachten_Netz.pdf?__blob=publicationFile&v=3

Facebook (2021a, January 27). *Facebook Reports Fourth Quarter and Full Year 2020 Results* [Press release]. Retrieved May 10 2021, from https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx

Facebook (2021b, August 18). *Community Standards Enforcement Report (Q2 2021)*. Facebook. Retrieved October 18 2021, from https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/

Fielitz, M., Schwarz, K. and Hitziger, J. (2020). *Hate Not Found?! Deplatforming the Far Right and Its Consequences*. Institut für Demokratie und Zivilgesellschaft, Antonio Amadeu Stiftung. Retrieved May 10 2021, from https://www.idz-jena.de/fileadmin//user_upload/Hate_not_found/IDZ_Research_Report_Hate_not_Found.pdf

Gagliardone, I. (2019). Defining Online Hate and Its 'Public Lives': What is the Place for 'Extreme Speech'? *International Journal of Communication*, *13*, 3068-86. https://ijoc.org/index.php/ijoc/article/view/9103

Gerstenfeld, P.B. (2018). *Hate Crimes: Causes, Controls, and Controversies* (4th ed.). London: Sage.

Gesche, D., Klaßen, A, Quent, M. and Richter, C. (2019). *#Hass Im Netz: Der Schleichende Angriff Auf Unsere Demokratie – Eine Bundesweite Repräsentative Untersuchung.* Institut für Demokratie und Zivilgesellschaft. Retrieved September 23 2021, from https://www.idz-jena.de/fileadmin/user_upload/_Hass_im_Netz_-_Der_schleichende_Angriff.pdf

Gillespie, T. and Aufderheide, P. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates – Introduction. *Internet Policy Review*, *9*(4), 2-4. https://doi.org/10.14763/2020.4.1512

Goldman, E. (2021). Content Moderation Remedies. *Michigan Technology Law Review*. Advance online publication. https://dx.doi.org/10.2139/ssrn.3810580

Gollatz, K. and Jenner, L. (2018). *Hate Speech und Fake News – Zwei verwobene und politisierte Konzepte*. Humboldt Institut für Internet und Gesellschaft. Retrieved May 10 2021, from https://www.hiig.de/hate-speech-fake-news-two-concepts-got-intertwined-politicised/

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

Gorwa, R. (2019b). What is platform governance? *Information, Communication & Society, 22*(6), 854-871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates – The Future of Regulating Content Moderation: Content Moderation Has a Regulatory Politics. *Internet Policy Review*, *9*(4), 18-19. https://doi.org/10.14763/2020.4.1512

Gorwa, R. (2021). Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG. *Telecommunications Policy, 45*(6). https://doi.org/10.1016/j.telpol.2021.102145

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). https://doi.org/10.1177/2053951719897945

Gorwa, R. and Garton Ash, T. (2020). Democratic Transparency in the Platform Society. In Persily, N. and Tucker, J.A. (Eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (1st ed., pp.286-312). Cambridge: Cambridge University Press.

Hagey, K. and Horwitz, J. (2021, September 15). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. *Wall Street Journal*. Retrieved September 20 2021, from https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Hao, K. (2021, March 11). How Facebook got addicted to spreading misinformation. *MIT Technology Review*. Retrieved May 10 2021, from https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Harbinja, E., Leiser, M.R., Barker, K., Mangan, D., Romero-Moreno, F. and Dushi, D. (2019). *BILETA Response to the UK Government Consultation 'Online Harms White Paper'*. British Irish Law Education and Technology Association. Retrieved May 10 2021, from https://uhra.herts.ac.uk/handle/2299/21431?show=full

Hartzog, W. and Selinger, E. (2015). Increasing the Transaction Costs of Harassment. *Boston University Law Review, 95*, 47-51.

Haufe (2021, April 1). Bundespräsident unterzeichnet Gesetz zur Bekämpfung von Hass und Rechtsextremismus. *Haufe Online Redaktion*. Retrieved May 10 2021, from https://www.haufe.de/recht/weitere-rechtsgebiete/strafrecht-oeffentl-recht/gesetzespaket-zur-bekaempfung-der-hasskriminalitaet-im-internet_204_510192.html

Haupt, C. (2021). Regulating Speech Online: Free Speech Values in Constitutional Frames. *Washington University Law Review*. Advance online publication. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3794884

He, D. (2020). Governing Hate Content Online: How the *Rechtsstaat* Shaped the Policy Discourse on the NetzDG in Germany. *International Journal of Communication, 14*, 3746-68. https://ijoc.org/index.php/ijoc/article/view/14213/3150

Heidrich, J. and Scheuch, B. (2017). Das Netzwerkdurchsetzungsgesetz: Anatomie eines gefährlichen Gesetzes. In: Taeger, J. (Ed.), *Recht 4.0: Innovationen aus den rechtlichen Laboren*, (1st ed., pp.305-319). Oldenburg: Oldenburger Verlag für Wirtschaft, Informatik und Recht.

Helberger, N. (2020). The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. *Digital Journalism*, *8*(6), 842-854. https://doi.org/10.1080/21670811.2020.1773888

Heldt, A.P. (2019a). Let's meet halfway: Sharing new responsibilities in a digital age. *Journal of Information Policy*, *9*, 336-369. https://doi.org/10.5325/jinfopoli.9.2019.0336

Heldt, A.P. (2019b). Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2), 336-369. https://doi.org/10.14763/2019.2.1398

Heldt, A.P. (2020). Intermediärsregulierung: Quo Vadis NetzDG & Co? *UFITA, 84*, 529-542.

Hern, A. (2021, August 18). TechScape: How the UK forced global shift in child safety policies. *Guardian*. Retrieved September 6 2021, from https://www.theguardian.com/technology/2021/aug/18/uk-governments-child-safety-regulation-leads-to-global-policy-shifts

Hiéramente, M. (2021, January 15). Groβe Koalition plant das nächste verfassungswidrige Gesetz. *Netzpolitik*. Retrieved May 10 2021, from https://netzpolitik.org/2021/bestandsdatenauskunft-grosse-koalition-plant-das-naechste-verfassungswidrige-gesetz/

Hirschl, R. (2005). The Question of Case Selection in Comparative Constitutional Law. *American Journal of Comparative Law, 53*(1), 125-155. https://doi.org/10.1093/ajcl/53.1.125

Horwitz, J. and Seetharaman, D. (2020, May 26). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. *The Wall Street Journal*. Retrieved May 10 2021, from

https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Hosseinmardi, H., Ghasemian, A., Clauset, A., Rothschild, D.M., Mobius, M. and Watts, D.J. (2020). *Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube*. Manuscript in preparation. Retrieved May 25 2021, from https://arxiv.org/pdf/2011.12843.pdf

Jacob, L. (2018, April 17). 87% of Germans Approve of Social Media Regulation Law. *Dalia Research*. Retrieved May 25 2021, from https://daliaresearch.com/blog/blog-germans-approve-of-social-media-regulation-law/

Johnson, N.F., Leahy, R., Restrepo, N.J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P. and Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261-265. https://doi.org/10.1038/s41586-019-1494-7

Kaye, D. (2017). *Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (OL DEU 1/2017). United Nations Office of the High Commissioner for Human Rights. Retrieved May 10 2021, from https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf

Kayser-Bril, N. (2020). *Automated moderation tool from Google rates People of Color and gays as 'toxic'*. AlgorithmWatch. Retrieved May 10 2021, from https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/

Keller, D. (2021). *Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List*. Center for Internet and Society at Stanford Law School. Retrieved May 10 2021, from https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws

Keum, B.T. and Miller, M.J., 2018. Racism on the Internet: Conceptualization and recommendations for research. *Psychology of Violence*, *8*(6), 782-91. https://doi.org/10.1037/vio0000201

Klausa, T. (2021, September 3). NetzDG-Verstöße: Facebook zahlt fünf Millionen Euro. *Tagesspiegel Background*. Retrieved September 6 2021, from https://background.tagesspiegel.de/digitalisierung/netzdg-verstoesse-facebook-zahlt-fuenf-millionen-euro

Klein, A. (2012). Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, *22*(4), 427-48. https://doi.org/10.1111/j.1468-2885.2012.01415.x

Klonick, K. (2018). The New Governors: The People, Rules and Processes Governing Online Speech. *Harvard Law Review, 131*, 1598-670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Koehler, D. (2018). Recent trends in German right-wing violence and terrorism: what are the contextual factors behind 'hive terrorism'? *Perspectives on Terrorism*, *12*(6), 72-88.

Kümpel, A.S. and Rieger, D. (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien Ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation*. Konrad-Adenauer-Stiftung. Retrieved May 10 2021, from https://epub.ub.uni-

muenchen.de/68880/1/Rieger_Wandel%20der%20Sprach-%20und%20Debattenkultur%20in%20sozialen%20Online-Medien.pdf

Lamensch, M. (2021, April 22). When Women Are Silenced Online, Democracy Suffers. *CIGI*. Retrieved May 10 2021, from https://www.cigionline.org/articles/when-women-are-silenced-online-democracy-suffers

Landesanstalt für Medien NRW (2018). *Ergebnisbericht Hassrede*. Retrieved May 10 2021, from https://www.medienanstalt-nrw.de/zum-nachlesen/forschung/abgeschlossene-projekte/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Leerssen, P. (2015). Cut Out by the Middle Man: The Free Speech Implications of Social Network Blocking and Banning in the EU. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 6*(2), 99-119. https://www.jipitec.eu/issues/jipitec-6-2-2015/4271

Leerssen, P., (2020). The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. *European Journal of Law and Technology, 11*(2). http://ejlt.org/index.php/ejlt/article/view/786

Lessig, L. (1999). *Code and Other Laws of Cyberspace*. New York: Basic Books.

Lewis, R. (2018). *Alternative influence: Broadcasting the reactionary right on YouTube*. Data & Society Research Institute. Retrieved May 10 2021, from https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

Lewis, R. (2020, January 8). *All of YouTube, Not Just the Algorithm, is a Far-Right Propaganda Machine*. *Medium*. Retrieved May 10 2021, from https://ffwd.medium.com/all-of-youtube-not-just-the-algorithm-is-a-far-right-propaganda-machine-29b07b12430

Liesching, M. (2020). *Stellungnahme zum Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes*. Deutscher Bundestag Ausschuss für Recht und Verbraucherschutz. Retrieved October 11 2021, from https://www.bundestag.de/resource/blob/700788/83b06f596a5e729ef69348849777b045/liesching-data.pdf

Maas, H. (2017). *Rede des Bundesministers der Justiz und für Verbraucherschutz, Heiko Maas, zum Gesetzentwurf zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz) vor dem Deutschen Bundestag am 30. Juni 2017 in Berlin: Bulletin 80-1*. Bundesregierung. Retrieved May 10 2021, from https://www.bundesregierung.de/breg-de/service/bulletin/rede-des-bundesministers-der-justiz-und-fuer-verbraucherschutz-heiko-maas--793138

Mac, R. and Silverman, C. (2020, December 11). After the US Election, Key People Are Leaving Facebook and Torching the Company in Departure Notes. *Buzzfeed News*. Retrieved May 10 2021, from https://www.buzzfeednews.com/article/ryanmac/facebook-rules-hate-speech-employees-leaving

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *new media & society*, *19*(3), 329-46. https://doi.org/10.1177/1461444815608807

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930-46. https://doi.org/10.1080/1369118X.2017.1293130

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S.K., Goyal, P. and Mukherje, A. (2019). Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media, 13*(1), 369-380. https://ojs.aaai.org//index.php/ICWSM/article/view/3237

Matias, J.N. (2016, December 12). *The Obligation to Experiment*. Medium. Retrieved May 10 2021, from https://medium.com/mit-media-lab/the-obligation-to-experiment-83092256c3e9

Mchangama, J. and Alkiviadou, N. (2020). The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act Two. Justitia. Retrieved May 10 2021, from https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf

Munger, K. and Phillips, J. (2020). Right-Wing YouTube: A Supply and Demand Perspective. *International Journal of Press/Politics*. https://doi.org/10.1177/1940161220964767.

Neuberger, C. (2018). Meinungsmacht im Internet aus kommunikationswissenschaftlicher Perspektive. *UFITA*, *82*(1), 53-68.

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

NTV (2018, 29 August). NetzDG erschwert Rechten Rekrutierung. *NTV*. Retrieved May 10 2021, from https://www.n-tv.de/politik/NetzDG-erschwert-Rechten-Rekrutierung-article20597308.html

Oberlandesgericht Braunschweig, *Urteil vom 5. Februar 2021* (1 U 9/20), https://openjur.de/u/2318910.html

Oltermann, P. (2018, January 5). Tough new German law puts tech firms and free speech in spotlight. *Guardian*. Retrieved May 10 2021, from https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight

Papaevangelou, C. (2021). The existential stakes of platform governance. *Open Research Europe, 1*(31). https://doi.org/10.12688/openreseurope.13358.2

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, Massachusetts: Harvard University Press.

Post, R. (2009). Hate Speech. In Hare, I. and Weinstein, J. (Eds.), *Extreme Speech and Democracy* (1st ed., pp.123-38). Oxford: Oxford University Press.

Rauchfleisch, A. and Kaiser, J. (2020). The German Far-Right on YouTube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, *64*(3), 373-96. https://doi.org/10.1080/08838151.2020.1799690

Reinemann, C., Niemierza, A., Fawzi, N., Riesmeyer, C. and Neumann, K. (2019). *Jugend-Media-Extremismus: Wo Jugendliche mit Extremismus in Kontakt kommen und wie sie ihn erkennen.* Wiesbaden: Springer.

Reuters (2021, July 27). Google takes legal action over Germany's expanded hate-speech law. *Reuters*. Retrieved September 6 2021, from https://www.reuters.com/technology/google-takes-legal-action-over-germanys-expanded-hate-speech-law-2021-07-27/

Rosen, G. (2020). *Recommendation Guidelines*. Facebook. Retrieved May 10 2021, from https://about.fb.com/news/2020/08/recommendation-guidelines/

Salter, M. and Mason, J. (2007). *Writing Law Dissertations: An Introduction and Guide to the Conduct of Legal Research.* London: Longman.

Schulz, W. (2018). *Regulating Intermediaries to Protect Privacy Online – the Case of the German NetzDG.* Humboldt Institut für Internet und Gesellschaft. Retrieved May 10 2021, from https://www.hiig.de/wp-content/uploads/2018/07/SSRN-id3216572.pdf

Seetharaman, D., Horwitz, J. and Scheck, J. (2021, October 17). Facebook Says A.I. Will Clean Up the Platform. Its Own Engineers Have Doubts. *Wall Street Journal.* Retrieved October 10 2021, from https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184

Sellars, A. (2016). *Defining hate speech*. Berkman Klein Center. Retrieved May 10 2021, from https://cyber.harvard.edu/publications/2016/DefiningHateSpeech

Siapera, E. and Viejo-Otero, P. (2021). Governing Hate: Facebook and Digital Racism. *Television and New Media, 22*(2), 112-130. https://doi.org/10.1177%2F1527476420982232

Siegel, A. (2020). Online Hate Speech. In Persily, N. and Tucker, J.A. (Eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform* (1st ed., pp.56-88). Cambridge: Cambridge University Press.

Siegel, A. and Badaan, V. (2020). #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837-855. https://doi.org/10.1017/S0003055420000283

Siems, M. (2009). The Taxonomy of Interdisciplinary Legal Research: Finding the Way Out of the Desert. *Journal of Commonwealth Law and Legal Education, 7*(1), 5-17. https://doi.org/10.1080/14760400903195090

Shephard, N. (2020). *Digitale Gewalt an Frauen: Was kann das NetzDG?* Heinrich Böll Stiftung, Gunda-Werner-Institut für Feminismus und Geschlechterdemokratie. Retrieved May 10 2021, from https://www.gwi-boell.de/de/2020/03/03/digitale-gewalt-frauen-was-kann-das-netzdg

Soral, W., Bilewicz, M. and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136-46. https://doi.org/10.1002/ab.21737

Stark, B. and Stegmann, D. (2020). *Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse.* AlgorithmWatch. Retrieved May 10 2021, from

https://algorithmwatch.org/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020-AlgorithmWatch.pdf

Stenner, P. and Reuter, M. (2021, July 9). Telegram soll sich an das NetzDG halten. *Netzpolitik*. Retrieved September 6 2021, from https://netzpolitik.org/2021/bussgeldverfahren-telegram-soll-sich-an-das-netzdg-halten/

Suzor, N., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J. and Van Geelen, T. (2019). Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. *Policy & Internet*, *11*(1), 84-103. https://doi.org/10.1002/poi3.185

Taekema, S. (2018). Theoretical and normative frameworks for legal research: Putting theory into practice. *Law and Method*. https://doi.org/10.5553/REM/.000031

Tworek, H.J.S. (2021). Fighting Hate with Speech Law: Media and German Visions of Democracy. *Journal of Holocaust Research, 35*(2), 106-122. https://doi.org/10.1080/25785648.2021.1899510

Tworek, H.J.S. and Leerssen, P. (2019). *An Analysis of Germany's NetzDG Law*. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Retrieved May 10 2021, from https://pure.uva.nl/ws/files/40293503/NetzDG_Tworek_Leerssen_April_2019.pdf

Udupa, S. and Pohjonen, M. (2019). Extreme Speech and Global Digital Cultures – Introduction. *International Journal of Communication*, *13*, 3049-67. https://ijoc.org/index.php/ijoc/article/view/9102

Vaidhyanathan, S. (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford: Oxford University Press.

Van Drunen, M.Z. (2020). The post-editorial control era: how EU media law matches platforms' organisational control with cooperative responsibility. *Journal of Media Law, 12*(2), 166-190. https://doi.org/10.1080/17577632.2020.1796067

Van Loo, R. (2020). The New Gatekeepers: Private Firms as Public Enforcers. *Virginia Law Review, 106*, 467-522. https://www.virginialawreview.org/wp-content/uploads/2020/12/VanLoo_Book.pdf

Wagner, B., Rozgonyi, K., Sekwenz, M.T., Cobbe, J. and Singh, J. (2020). Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30,* 261-271. https://doi.org/10.1145/3351095.3372856

Wegrich, K. and Hammerschmid, G. (2018). *Public administration characteristics and performance in EU28: Germany*. European Commission Directorate-General for Employment, Social Affairs and Inclusion. Retrieved May 10 2021, from https://op.europa.eu/en/publication-detail/-/publication/0f22ae85-9619-11e8-8bc1-01aa75ed71a1/language-en

Whittaker, J., Looney, S., Reed, A. and Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review, 10*(2). https://doi.org/10.14763/2021.2.1565

Wischmeyer, T. (2020). What is illegal offline is also illegal online: The German Network Enforcement Act 2017. In Petkova, B. and Ojanen, T. (Eds.), *Fundamental Rights Protection Online* (1st ed., pp.26-56). Cheltenham: Edward Elgar Publishing.

York, J.C. (2021). *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. New York: Verso.

YouTube (2019). *The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation*. YouTube Official Blog. Retrieved May 10 2021, from https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce

YouTube (2020). *Help us keep comments respectful – new Community Guidelines comment reminders*. YouTube Help. Retrieved May 10 2021, from https://support.google.com/youtube/thread/86685658

Appendix I:
List of Interviewees

Pseudonym	Role/s	Interview date
Interviewee A	Board member of an association for feminist digital policy Responsible for gender equality and digital policy at a political foundation	23 March 2021
Interviewee B	Complaints officer at a *Beschwerdestelle*[14] run by a technology industry association[15]	19 March 2021
Interviewee C	Board member of an association for feminist digital policy Deputy director of a research centre for technology and human rights Formerly worked for an anti-racist NGO and for a non-profit digital media foundation	30 March 2021
Interviewee D	Head of digital policy for an anti-racist NGO Journalist and editor focusing on racism and right-wing extremism	13 April 2021
Interviewee E	Police officer Volunteer at an NGO which runs a *Beschwerdestelle*[15] and campaigns against (especially homophobic) online hate speech	12 March 2021
Interviewee F	Independent consultant focusing on online hate speech Project leader for an anti-hate speech campaign run by an association for diversity and equality in journalism	19 March 2021
Interviewee G	Political scientist and policy researcher Advisor to multiple EU projects focusing on extremism and digital democracy	25 March 2021

All interviews were conducted in English, with the exception of Interviewee E, who was interviewed in German. Quotes from Interviewee E have been translated by the author.

Interviewees A-E were identified using the list of German anti-hate speech initiatives maintained by Das Nettz (n.d.). Interviewee F was suggested by a personal contact of the author and Interviewee G was suggested by Interviewee F.

[14] A non-governmental organisation which provides an alternative channel for internet users to submit complaints about illegal content. The *Beschwerdestelle* then follows up with platforms, server operators etc. in order to get the content removed, and in some cases with the police. There are a number of such organisations operating in Germany.
[15] Four platforms which are subject to NetzDG are members of the association in question.

Appendix II:
Standard Interview Questionnaire

- Could you tell me about your opinion of NetzDG generally and the position you would take towards the law?

- In your opinion, what have been the biggest effects of NetzDG since it came into force?

    - Has it affected the probability that people encounter hate speech on social media?

    - Has it affected the probability that people share hate speech on social media?

- In light of your experiences working in this area, how would you describe the experiences of people who are affected by online hate speech?

    - Have these experiences changed at all since NetzDG was introduced? If so, how?

    - How would you describe the experience of reporting hate speech to platforms under NetzDG?

    - How helpful do you think this reporting option is for victims?

    - Are there any improvements to this process you would recommend?

- In your opinion, how have the big social media platforms changed their practices since NetzDG was introduced? (If at all)

- Can you tell me about your opinion on the two reform proposals of NetzDG that were launched last year?

- What would be your most important recommendations for the federal government on how they could better address online hate speech? For platforms?

Appendix III:
Illegal Content

§1(3) NetzDG provides that:

> 'Unlawful content shall be content within the meaning of §1(1) [content disseminated on social networks] which fulfils the requirements of the offences described in §§86, 86a, 89a, 91, 100a, 111, 126, 129 to 129b, 130, 131, 140, 166, 184b, 185 to 187, 201a, 241 or 269 of the Criminal Code and is not justified.'[16]

As outlined in section 2(a), for the purposes of this paper, 'hate speech' is taken to describe any criminal speech act motivated by prejudice against one of the groups listed in the BKA's definition of a hate crime. Any of these offences could therefore qualify as hate speech if committed with such motivations. However, some of the specified offences (e.g. §130, §201a, §241) are more common in the social media context, and more likely to be motivated by such prejudices, than others (e.g. §89a, §100a, §269).

StGB provision	Description
§86	Dissemination of propaganda material of unconstitutional organisations (e.g. banned political parties, Nazi propaganda)
§86a	Use of symbols of unconstitutional organisations (e.g. flags, uniforms, slogans)
§89a	Preparation of serious violent offences endangering the state (e.g. by instructing another person or receiving instruction in the use of firearms, explosives or other weapons)
§91	Instructions for committing a serious violent offence endangering the state
§100a	Treasonous forgery (forgery which could deceive a foreign power such as to seriously damage Germany's external security or foreign relations)
§111	Public incitement to commit a crime
§126	Disturbing the public peace by threatening to commit a crime (which must be a serious offence such as murder or grievous bodily harm) or by knowingly pretending that the commission of such a crime is imminent
§129	Forming a criminal organisation
§129a	Forming a terrorist organisation
§129b	[not an offence; specifies procedural rules relating to §§129 and 129a]
§130	Incitement of the masses (this includes inciting hatred against a national, racial, religious or ethnic group; calls for violence against such a group; insults violating human dignity; dissemination of content inciting hatred or violence; denying the crimes of the Nazi regime; and glorification of Nazism)
§131	Depictions of violence (in a manner which glorifies or downplays the violence or otherwise violates human dignity)
§140	Rewarding or publicly approving certain serious offences
§166	Revilement of religious faiths and religious and ideological communities (in a manner which could disturb the public peace)
§184b	Dissemination, procurement and possession of child sexual abuse material
§185	Insult
§186	Malicious gossip
§187	Defamation
§201a	Violation of intimate privacy by taking photographs or other images

[16] BMJV official translation (BMJV, 2021).

§241	Threatening to commit a serious criminal offence
§269	Forgery of data of probative value (for the purposes of deception in legal commerce)