Heldt, Amélie

Article

Reading between the lines and the numbers: an analysis of the first NetzDG reports

Internet Policy Review

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

Suggested Citation: Heldt, Amélie (2019) : Reading between the lines and the numbers: an analysis of the first NetzDG reports, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 8, Iss. 2, pp. 1-18, https://doi.org/10.14763/2019.2.1398

This Version is available at:
https://hdl.handle.net/10419/214071

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 8 | Issue 2

Reading between the lines and the numbers: an analysis of the first NetzDG reports

Amélie Heldt
Leibniz Institut für Medienforschung, Hans-Bredow-Institut, Hamburg, Germany

Published on 12 Jun 2019 | DOI: 10.14763/2019.2.1398

Abstract: Approaches to regulating social media platforms and the way they moderate content has been an ongoing debate within legal and social scholarship for some time now. European policy makers have been asking for faster and more effective responses from the various social media platforms to explain how they might deal with the dissemination of hate speech and disinformation. After a failed attempt to push social media platforms to self-regulate, Germany adopted a law called the Network Enforcement Act (NetzDG) which forces platforms to ensure that "obviously unlawful content" is deleted within 24 hours. It contains an obligation that all platforms that receive more than 100 complaints per calendar year about unlawful content must publish bi-annual reports on their activities. This provision is designed to provide clarification on the way content is moderated and complaints handled on social networks. After the NetzDG came into force, initial reports reveal the law's weak points, predominantly in reference to their low informative value. When it comes to important takeaways regarding new regulation against hate speech and more channelled content moderation, the reports do not live up to the expectations of German lawmakers. This paper analyses the legislative reasoning behind the reporting obligation, the main outcomes of the reports from the major social networks (Facebook, YouTube, and Twitter) and why the reports are unsuitable to serve as grounds for further development of the NetzDG or any similar regulation.

Keywords: NetzDG, Speech regulation, Social media, Transparency, Content moderation

Article information

Received: 29 Aug 2018 **Reviewed:** 31 May 2019 **Published:** 12 Jun 2019

Competing interests: The author has declared that no competing interests exist that have influenced the text.

URL:
http://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports

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

INTRODUCTION

Good content, bad content? What seems reasonable to some might be offensive to others. Depending on our social norms, laws and culture, we tend to categorise what we see on the internet as a process of content selection that fits our expectations and our needs or not (Gillespie, 2018, p. 197). This plays into our perceptions of the content disseminated by users on social media platforms and these inherent discrepancies constitute a reason why regulating online speech is still an unresolved issue for lawmakers. In the legislative process one pervading question remains: how to improve content moderation in the light of long established legal provisions. The minimum European lawmakers tend to agree upon is the legitimacy of the need to take down unlawful content. At the very least, this is what the German government assumed when it issued a law that made it mandatory for the largest social media platforms to ban obviously unlawful content within 24 hours. In doing so, Germany was one of the first countries to issue a so-called anti-"hate speech" law and it has been the target of fairly constant criticism since it was instated. Not only do critics label it as unconstitutional (Gersdorf, 2017; Schulz, 2018), but it is also mentioned in the international discussion as a bad example of platform regulation (Special Rapporteur for the UN: Kaye, 2018). The NetzDG was meant to enhance the protection of users against hate speech and to provide more clarity on the way platforms handle and moderate unlawful content (Wischmeyer, 2018, p. 7). However, as this paper will show, so far there is still no certainty about either of these goals mainly because the reports do not provide well defined results. Many have expected the published reports to provide more substantial insights on content moderation policies – an expectation that seems to have been betrayed (Gollatz et al., 2018).

The main aim of this article is to provide an analysis of the reports published by the major social media platforms, including insights into the implementation of the NetzDG while focusing on the obligation of ensuring user-friendly complaint procedures. The result of this analysis shows that it is not enough to require companies to publish transparency reports if the information they contain has no real informative value. This article shows that the law might have incentivised the platforms to remove hate speech faster, but that there is no certainty about its effect due to the lack of substantial information in the reports. Facebook for example does not fully comply with the obligation of supplying an easily recognisable complaint procedure, therefore the number of complaints cannot be considered conclusive. After introducing the NetzDG in general, and the reasons why it is deemed unconstitutional in the German debate, the present article dives deeper into the rationale of the NetzDG. This law can serve as an example of "new-school speech regulation", that is, a type of regulation that is aimed at the owners of the digital infrastructure instead of the speakers themselves (Balkin, 2014, p. 2298). Starting from the published complaint figures, this article examines the implementation of complaint tools and eventually goes to show that the reports constitute no factual ground for a re-evaluation of the NetzDG or for a similar regulatory project, even though the reporting obligation is a key provision. The law leaves room for interpretation with regards to its implementation and this has somehow hollowed out the obligation to publish transparency reports as the figures do not reflect the full picture. The reporting obligation under NetzDG might however serve as a counterexample in the discussion on transparency and the corresponding reports. It shows that we need to formulate transparency rules in a clearer way so that the data collected can serve the purpose of iteration for both the companies and the state.

1. THE NETZDG: AN ACT TO IMPROVE LAW ENFORCEMENT ON SOCIAL MEDIA PLATFORMS

1.1. THE REGULATORY RATIONALE OF THE NETZDG

A series of events involving online discrimination against and agitation towards ethnic minorities (that reached a peak in 2015-2016 when a wave of refugees from Syria arrived in Germany) triggered the German government in acting against hateful online content. There are other factors that could also be interpreted as catalysts that caused the government to act: an increasing awareness of the problem of aggressive and potentially harmful online communication and an air of mistrust of the tech companies that run the biggest social media platforms. .The latter especially applies to their governance through content moderation policies (Citron, 2017, p. 1065 f.). In order to combat hate speech and other unwanted content on social networks (cf. Delort et al., 2011, p. 9), the German government drafted a law that would force social networks to examine complaints and delete content if possible within 24 hours of receiving user complaints. The main motive for the legislator's action was to thwart the increase of hate speech on social networks (as stated in the law's explanatory memorandum) and to respond to public pressure surrounding the issue (Liesching, 2018c, para. 2). After several failed attempts to implement a system of self-regulation by the social network companies to reduce the proliferation of hate speech, the German Ministry of Justice drafted the NetzDG and it was finally ratified by Parliament in July 2017. The EU, for instance, began a self-regulatory initiative with the 2016 EU Code of Conduct on Countering Illegal Hate Speech Online in cooperation with Facebook, Microsoft, Twitter and YouTube (its fourth evaluation was published in February 2019).

The "Act to Improve Enforcement of the Law in Social Networks" (Network Enforcement Act, hereinafter NetzDG) became fully effective on 1 January 2018. It aims for the faster enforcement of German criminal laws, hence the deletion – when appropriate – of unlawful content. The law defines social networks as follows: "telemedia service providers which, for profit-making purposes, operate internet platforms which are designed to enable users to share any content with other users or to make such content available to the public" (official translation by German Ministry of Justice). Hate speech as such is – on the contrary – neither defined by the NetzDG nor by the German Penal Code (hereinafter StGB), although the general discussion around the legislative project mostly refers to the term. The general definition of hate speech is: speech designed to promote hatred on the basis of race, religion, ethnicity, national origin or other specific group characteristics (Rosenfeld, 2002, p. 1523; Djuric et al., 2015). The NetzDG refers to StGB sections without defining the offences listed. In the StGB these offences are not listed under one specific title but in different categories, for example, breaches of public order or libel. This creates a chain of provisions that refer to one another (see below). The offences targeted by the NetzDG can be conflated as hate speech when discussing the type of offences, but it is not actually a technical term under German law.

Although scholars agree on a general definition of hate speech, its criminal prosecution differs from one country to another (even within member states of the European Union). It is important to note, therefore, that no new criminal offences for online hate speech were created or added to the StGB. Instead, the NetzDG lists 22 offences that were already and still are punishable under the StGB, such as libel, defamation, sedition and calls for violence and adds a *de facto* enforcement obligation for large social media platforms. Anyone breaching these laws by posting, commenting or uploading content on social media platforms stills incurs a penalty

from the state. In addition, German law now forces social networks to become more active. They are obliged to implement procedures that ensure obviously unlawful content is deleted within 24 hours of receiving a complaint. If there is any doubt regarding a takedown decision, the procedure may take up to seven days. After that deadline, a final decision on the lawfulness of a post must be reached and unlawful content needs to be removed, that is, either blocked or deleted. The fines for a breach of this obligation can reach up to €50 million. In addition to complying with this operational provision, social media platforms are obliged to publish bi-annual reports which will be addressed and analysed below.

1.2. MAIN ALLEGATIONS REGARDING THE VIOLATION OF CONSTITUTIONAL LAW

In order to fully assimilate the importance of the published reports, it is helpful to know more about the context of the NetzDG. This law has been under attack ever since its first draft was made public. Not only was it perceived as an *ad hoc* legislative reaction, more importantly it is considered merely as a loophole that transfers public responsibility to a private actor. It has been criticised from many perspectives and an exhaustive description would go beyond the scope of this article (cf. Schulz, 2018, passim). However, a brief overview is necessary to comprehend the context in which the first NetzDG reports were published. To state it briefly, a wide array of scholars, politicians and activists demanded the abrogation of the NetzDG or a revised version in the near future. Liberal politicians have opposed the law in court and there have been counter-proposals (e.g., the Green Party's proposal from January 2019). The criticism did not diminish over the course of 2018 although the consequences for free speech were not as severe as expected, instead the numbers below show that the NetzDG did not really have an impact on content moderation. Nonetheless, a law has no legitimation to stay in effect if it is deemed unconstitutional, which is why it is still expected to be revised. The following passage will give a general overview of the allegations made against the NetzDG. The procedural points of criticism will be omitted because they are more technical and inherent to the German legal system and add little to the present argument. Others are related to human rights infringements and partly transferrable to similar law projects in other jurisdictions (cf. Funke, 2018) and are, therefore, more relevant to this paper. [1]

The main focus of the (non-procedural) criticism relates to possible violations of freedom of speech in various ways. The *de facto* obligation for social networks to delete manifestly unlawful content within 24 hours has raised questions pertaining to the potential overblocking of content and to the privatisation of the judiciary due to the interpretation and application of criminal law by private companies. These two elements combined can in turn have chilling effects on freedom of speech and we will take a closer look at them below. First, however, one must bear in mind that the question of whether the content targeted by NetzDG may be deleted is not my main concern. First, because such content is illegal and second, because takedown is still the most effective tool social media platforms make use of when it comes to hate speech (Citron & Norton, 2011, p. 1468; Klonick, 2018, p. 12). The main source of scepticism is the shift of responsibility towards private companies as a corollary of the obligations that have been laid upon them (Guggenberger, 2017b, p. 2582). In sum, scholars agree that the NetzDG is not really an exemplar of methods for fighting hate speech online (Gersdorf, 2017, p. 447; Guggenberger, 2017a, p. 101; Balkin, 2018, p. 28).

The German Basic Law allows lawmakers to restrict fundamental rights under certain conditions and freedom of speech may be constrained by general laws according to art. 5 (2) Basic Law. This includes criminalising offensive speech and it can have a horizontal effect between private parties when private actors require one another to observe the law. This is

acceptable as long as it does not result in overblocking. Overblocking is the term used when content is deleted or blocked for no substantial reason, because the incentive to immediately delete rather than perform more fundamental checks arises (Richter, 2017; Holznagel, 2018, p. 369). Criminal offences related to the protection of honour (such as libel) mostly overlap with the categories used by social networks in their "community guidelines". However, this overlap is not preserved when it comes to the specific elements of a criminal offence. One would need to know and practice (national) criminal law and consider the context of the generated content (Wimmers & Heymann, 2017, p. 100). These various parameters make it difficult to parse unlawful content in a short timeframe and that is what the social media platforms have been insisting upon in recent years when justifying the slow removal of hate speech. The problem lies in the risk that there is just not enough time to make accurate takedown decisions coupled with the high level of pressure of being fined. In this scenario, the net result could potentially be overblocking (Kaye, 2018, p. 7). In accordance with section 3 (2) NetzDG, social networks must provide a procedure that ensures the deletion of obviously unlawful content within 24 hours after a user complaint. If they fail to do so, as mentioned earlier, they risk a fine of up to €50 million, which makes the incentive for decisions in favour of takedowns stronger than before the implementation of the NetzDG (Schiff, 2018, p. 370). However, the fear of overblocking doesn't seem to have materialised when looking at the takedown numbers in the reports published by the companies concerned by the NetzDG (Wieschmeyer, 2018, p. 20).

However, the critique of a substantial shift of responsibility from the judiciary to platforms themselves is still under discussion. This is mainly due to the wording of sec. 3 (2) Nr. 2 NetzDG, that is, to delete "content that is manifestly unlawful". Generally, content-related regulation has to be as neutral as possible with regards to the opinions expressed, i.e., it is subject to a strict proportionality test (so-called "*Wechselwirkung*"). The scope of application of a content targeting law must be sufficiently precise to avoid too much room for interpretation, which could, in the case of the NetzDG, result in an overly broad definition of legal terms and an unsubstantial removal of content. Using an unspecified legal term such as "manifestly", although the law is applied by a private party and not by a judge, puts the power of the judiciary in its interpretation and the application of the law at risk. Clear legal definitions and specific criteria are necessary to constrain the platforms' discretion (Nolte, 2017, p. 556-558; Liesching, 2018a, p. 27; Wieschmeyer, 2018, p. 15-16; Belli, Francisco, & Zingales, 2017, p. 52; Nunziato, 2014, p. 10); leaving the interpretation of a key term of the bill too unspecified is considered as unconstitutional (FCC: BVerfGE 47, 109, 121; Wimmers & Heymann, 2017, p. 97-98; Reuter, 2018, p. 86) because this kind of interpretation is actually a core function of the judiciary.

Unspecified legal terms are usually only filled with meaning by court rulings. Up until then, they can be loaded with interpretations which might be revised later. In order to decide whether a statement is still within the boundaries of the law and therefore still protected by freedom of speech, a judge will have to examine the requirements mentioned above and potentially balance the fundamental rights of both parties. The result can then later subsequently be applied by private parties as a standard or a guideline. What is "manifestly" illegal? The NetzDG's explanatory memorandum defines it as follows: "Content is manifestly unlawful if no in-depth examination is necessary to establish the unlawfulness within the meaning of sec. 1 (3)." This sentence does not explain from which starting point an examination is considered "in depth", it leaves the original question of how to define "manifestly" unanswered. Still, users' right to take the platforms' decisions to court remains intact when social networks delete user-generated content under NetzDG. This possibility makes it unlikely that the NetzDG will be abrogated on the grounds of the 'privatisation' argument. Nevertheless, the complexity of this assignment, deciding whether content is unwanted but perhaps not unlawful, is a core element of the public

debate around content moderation (Kaye, 2018, p. 4). Because of the important implications for users' freedom of speech, one cannot help but wonder about the fact that the German lawmakers delegated this task to social media platforms instead of enhancing their own law enforcement forces (Schmitz & Berndt, 2018, p. 7; Wieschmeyer, 2018, p. 15-16; Buermeyer, 2017).

1.3. THE OBLIGATION TO IMPLEMENT A COMPLAINT PROCEDURE

All in all, German lawmakers were aiming for a faster response from social networks as to when content is reported as unlawful. It resulted in the obligation to ensure a procedure that would guarantee a reaction on "manifestly unlawful content" within 24 hours after receipt of the complaint. As mentioned above, this provision is one of the most criticised because of the uncertain effects it could possibly have on free speech (Keller, 2018, p. 2). This topic is rightly at the centre of the debate because scholars are only beginning to know more about the effects of these rules on the behaviour of both the platforms and the users (Gollatz et al., 2018). However, this paper focuses more on the way social networks have implemented the obligation to ensure a complaint procedure. This will later contribute to measure the informational value of the reports (see infra, section 4).

According to sec. 3 (1) NetzDG, social networks have "to supply users with an easily recognisable, directly accessible and permanently available procedure for submitting complaints about unlawful content." The implementation of this obligation is decisive for the relevance of the reports when it comes to conducting meaningful evaluation and regulatory impact assessments. I evaluated the significance of the reports in view of the accessibility of the complaint tool for users and their comparativeness. As mentioned above, the number of complaints filed could be of significance as far as the regulatory goals are concerned if the manner in which the data was collected and presented in the reports was different, or on the other hand, if the provision was identically implemented, regardless of the company carrying out their legal obligations. One needs to bear in mind (again) the explanatory memorandum of the NetzDG which states that social networks have to provide a "user-friendly procedure for submitting complaints about unlawful content". Furthermore, the procedure must be "easily recognisable, immediately accessible and always available". The memorandum does not provide any further expectations or provisions concerning the implementation of the complaint tool. The second half of the memorandum's section on the complaint procedure contains the requirements for the way that complaints are handled once submitted by a user. It does not elaborate any further on the way social networks should design the complaint procedure as such, leaving the concrete implementation of complaint procedures, for the most part, at the platform's discretion. This aspect was probably not expected to be as decisive as it appears to be now that the reports show that at least one platform violates this provision. The criteria mentioned above regarding expectations of the "user-friendly procedure" shall therefore be at the centre of the remarks below when it comes to the informative value of the reports.

2. THE NETZDG REPORTING OBLIGATION

In order to gain a better understanding of the way social networks moderate user-generated content and how they decide whether or not to remove content, the German lawmakers included a biannual reporting obligation. According to section 2 NetzDG:

> Providers of social networks which receive more than 100 complaints per calendar year about unlawful content shall be obliged to produce half-yearly German-language

> reports on the handling of complaints about unlawful content on their platforms, covering the points enumerated in subsection (2), and shall be obliged to publish these reports in the Federal Gazette and on their own website no later than one month after the half-year concerned has ended. The reports published on their own website shall be easily recognisable, directly accessible and permanently available.[2]

The NetzDG's explanatory memorandum states that the reporting obligation is required "in order to create the necessary transparency for the general public", [3] a requirement that was similarly formulated a while ago by scholars and activists (Gillespie, 2018, p. 199). The secondary goal of the reporting obligation is to provide numbers and facts "necessary in the interest of an effective impact assessment". The German Parliament is currently discussing a revision of the law subsequent to proposals that range from a complete abrogation of the law to only light adaptations. Changes made to the NetzDG will be based, at least partly, on the reports. The whole NetzDG project also serves as an example (for better or worse) for similar legislative undertakings. These reports are, therefore, central to a better development of the legislative tool, not only at national level, but also to answer the challenge posed by content moderation in general. As mentioned above, the German approach was quite a push forward due to political circumstances and public pressure. There is so far no equivalent in other jurisdictions and no solutions considered standard nor has best practice been established across borders because, when it comes to balancing content moderation and freedom of expression, the issues that arise are too numerous and too varied.

Hate speech, fake news, copyright infringements – just to name a few of the issues that arise – are often confused in the public debate and their respective definitions differ from one country to another. Considering the fact that social media platforms act globally, a one-size-fits-all solution would reduce costs. At the same time, such an extensive approach could be a threat to freedom of expression because of scopes of application that are too broad, leading to more restrictive regimes. To design a new regulatory framework, it is, therefore, necessary to monitor the effectiveness of its application. On that account, implementing a reporting obligation in sec. 2 NetzDG was necessary to improve this type of regulation (Eifert, 2017, p. 1453). The memorandum also states that producing the reports shall "ensure a meaningful and comprehensive picture of how they [the social networks] deal with complaints". As sec. 2 (2) NetzDG determines the minimum requirements for the reports, the memorandum justifies the reporting obligation with the special role of social networks. They are "crucial to the public debate" and must take on their "increased social responsibility". Rather than providing numbers without context, the reports are supposed to help understand the connection between the grounds on which social networks delete or block unlawful content and the provisions provided by law. Unfortunately, this expectation was not fulfilled.

The minimum requirements for the reports include specific points under sec. 2 NetzDG, such as providing the "number of incoming complaints about unlawful content" (nr. 3), the "number of complaints for which an external body was consulted" (nr. 6) and the "number of complaints in the reporting period that resulted in the deletion or blocking of the content at issue" (nr. 7). The numbers listed in nr. 7 need to be broken down according to the reasons for the specific complaint which makes them particularly interesting with regards to the regulatory goal. The explanatory memorandum of the NetzDG is quite brief on that point: it merely mentions "the interests of transparency and the effectiveness of the complaint management" as the reason for that specific point and then refers to the comments on sec. 3 NetzDG ("Handling of complaints about unlawful content"). This part nevertheless reveals the tight connection between the

reports and the handling of complaints. As a result, the obligation to report mainly serves to enhance transparency which goes hand in hand with an effective impact assessment of the new law, as stated in the memorandum, and the long-term goal of developing this regulatory framework in a sensible manner. These goals are important characteristics for the evaluation of the published reports. It will become clear at a later point in this article that they were perhaps underrated and minimised by the platforms – as propositioned by the title.

3. MAIN RESULTS OF THE FIRST ROUND OF REPORTS

According to sec. 1 (1) NetzDG, only social networks that have more than two million users have to comply with its rules and, therefore, with the reporting obligation in sec. 2. In view of the user numbers on the largest social networks, the minimum of 100 complaints per calendar year (as an obligation for having to publish reports) was easily arrived at by Facebook, YouTube and Twitter. Their reports demonstrate many similarities in the way they handle the matter of content moderation, but they also feature notable differences as far as the numbers of complaints are concerned. [4] The three reports from Facebook, YouTube and Twitter were analysed for this article not long after their publication, in August 2018. The overall result, as will be explained below, is that provisions for these types of reports need to be precise if one wishes to gather meaningful data. In substance, the reports show that social media platforms tend to moderate content on the grounds of their own community guidelines more than on the basis of national criminal law. I presume that the reason for this is that it allows them to react on a global scale rather than on a national one. Furthermore, social media platforms tend to use terms and tonalities in their community guidelines that are very similar to the vocabulary used in the NetzDG, making it rather unclear to the user where the differences lie (cf. Celeste, 2018). This similarity between reports being stated, the divergence between the complaint figures is quite significant.

3.1. COMMUNITY GUIDELINES ARE PRIORITISED

As the reports show, the content review process is based on a two-step approach for all three platforms. After being notified of a user complaint, the first check is made on the grounds of community guidelines or standards (both terms being used synonymously hereinafter). If the content violates these internal rules the reviewer will take the content down. Only if the result of that review is negative and if the user also submitted a complaint under NetzDG provisions (not only community guidelines), the content will be further checked for NetzDG infringements. It remains unclear how much content was taken down as hate speech under community guidelines, which could also have been blocked because of a violation of German criminal law. To submit a complaint under the NetzDG, the user will either have to tick an additional NetzDG box in the case of YouTube and Twitter, or, in the case of Facebook, go to the "Help Centre" and follow a specific NetzDG complaint link. In the next subsection, I will take a closer look at how each platform implemented the NetzDG complaint procedure and the subsequent effects on their complaint numbers. The reports do not state whether complaints have been examined on NetzDG violations even if they were not flagged as such by users. Nevertheless, it appears that YouTube, Twitter and Facebook all prioritise their own community guidelines since none of them offers to immediately submit a complaint under NetzDG (which is not mandatory under sec. 3 (2) NetzDG).

A reason for this prioritisation could be the subsequent takedown options. So-called unwanted content, that is, content that violates community guidelines, will be deleted globally whereas

content that is unlawful under German penal law (and therefore subject to removal under NetzDG) could only be blocked in Germany. Considering that content might be illegal in several countries, deleting it according to community guidelines might be more effective than taking it down for just one single country, with the possibility of repeating this action in another country further down the line. This raises questions of freedom of expression in privately-owned communication spaces, especially with regards to collateral censorship (Eifert, 2017, p. 1452; Balkin, 2018, p. 6). Although, from a European perspective there is a big overlap between *unlawful* and *unwanted* content, the definitions do not completely intersect. From a communications science perspective, it might be questionable as to which consequences the NetzDG provisions have on the way social networks formulate their community guidelines on a global scale because they could adjust their own policies to fit the broadest definition of hate speech (Gollatz et al., 2018, p. 6). It also points out that adapting the community guidelines to national (criminal) law in order to decrease the differences between community rules and German law could, in turn, have a massive influence on another country's regulation of social networks. However, there is no certainty that platforms will follow that path because the cost of adapting to national legislation could be too high. Instead, they could broaden their community guidelines to cover multiple definitions of hate speech – eventually restricting more speech than necessary under respective national regulations (Eifert, 2017, p. 1452; Belli, Francisco, & Zingales, 2017, p. 46; Koebler & Cox, 2018).

3.2. CONTENT MODERATION: HOW?

Under the reporting obligation of sec. 2 NetzDG, it is mandatory to describe the personnel and organisational resources deployed to comply with the provisions of the law. The reports show that human content reviewers are by no means replaced by machines (Delort et al., 2011, p. 24). None of the three social networks examined for this paper solely rely on algorithms or artificial intelligence to cover the tasks of recognising and reviewing unlawful content or handling user complaints. Given the amount of data uploaded by users, filters and other, technologies are in use and undergo constant optimisation. Yet, in order to properly review content that might be unlawful, platforms still heavily depend on human moderators (Roberts, 2016; Matsakis, 2018). The role of moderators in the review process is even more important when it comes to evaluating content that does not violate community standards but is potentially unlawful (see infra, section 3.1.). Cases that do not violate community guidelines but might be punishable under German criminal law are in general more complex and as a result require more sophisticated review. The cost of content moderation therefore increases when it has to comply with the law rather than with community guidelines, this is because of the complexity of applying legal provisions instead of internal guidelines. From a cost-benefit point of view, platforms would prefer to minimise staff overheads (Read, 2019) which is why the way they address this challenge is worth paying attention to.

Through a partnership with a German company named Arvato, Facebook has a team specially dedicated to NetzDG complaints numbering approximately 65 employees (as of August 2018). The staff speaks German and is trained to handle the complaints that are within the scope of section 1 (3) NetzDG. As an initial step, the "Community Operations" team reviews the reported content to determine whether or not it violates Facebook's Community Standards. If the issue is taken further, the second part of the two-step approach, a so-called "Legal Takedown Operation" team takes over, who are specially trained to review content for potential illegality. The report does not mention any software that would support the work of the Facebook reviewers. This might indeed not be necessary since the amount of complaints under NetzDG seems quite manageable (see infra, section 3.3). In general, Facebook makes use of AI to identify content that clearly violates their community standards (Koebler & Cox, 2018), but relies on

approximately 15,000 moderators worldwide to review unwanted content (Newton, 2019) as hate speech is still difficult to identify automatically with any precision (Koebler & Cox, 2018).

YouTube has integrated tools such as content filters in their upload process. Their NetzDG report states that they already filter any video uploaded for unlawful content and they interpret this measure as an additional compliance with the provisions of sec. 1 (3) NetzDG. YouTube had to deal with copyright infringements a long time before the hate speech problem became so virulent and has therefore been using their filtering software, ContentID to manage copyright issues. Since June 2017, YouTube has also integrated machine learning in order to optimise the filtering of illegal content (YouTube, 2017). Furthermore, there is a team dedicated to NetzDG flagged content that, similar to Facebook's approach, does not violate community guidelines but is reported by a user or a complaints body as a NetzDG violation. For the time being, their team dedicated solely to NetzDG complaints numbers approximately 100 employees.

According to Twitter's report, over 50 people work on NetzDG complaints. The report does not include any further information on the technological tools used to support that team in any way. Given Twitter's massive deletion of (presumably) fake accounts in July 2018 on the one hand, and the immense volume of content constantly uploaded on the other, it is quite likely that Twitter also uses filters to detect unwanted content. The company does not disclose which technological resources it uses to find and review unwanted content such as hate speech. However, it probably makes use of algorithms and machine learning to support human reviewers to reduce the amount of labour involved in the process. Such filters could also be used by Twitter to proactively detect content that is potentially unlawful under NetzDG but it does not mention them in its report (this information is not mandatory under NetzDG reporting provisions).

Since automated technologies are not yet able to detect cases related to sensitive or context-related issues that can be classified as as hate speech, satire, awareness- or education-related or even politically sensitive, they are unable to handle all types of complaints. Although studies show that the technology is getting better at detecting hate speech and offensive language using machine learning (Gaydhani et al., 2018), researchers tend to agree that so far no technology is capable of recognising hate speech beyond very clear cases of offensive speech (Gröndahl et al., 2018, p. 3). YouTube's statement on this subject is clear and straightforward: "Algorithms are not capable of recognising the difference between terror propaganda and critical coverage of such organisations or between inciting content and political satire." The inability of filters to recognise unwanted content could be the cause of unsubstantial content removal. Several cases were left uncommented on by the concerned platforms, including satire and innocent pictures, which could be because of human mistakes but perhaps also because of algorithmic and intelligent systems' failure to distinguish unwanted content from uncontentious content. These cases have been discussed in the media, but none of the platforms disclosed on which grounds the content was removed (Schmitz & Berndt, 2018, p. 31). Taken together, the point on the "resources deployed to comply with the NetzDG" constitutes only one of many sources of speculation around the use of technology in content moderation and the reports are not specific enough to draw further conclusions in that area.

3.3. NUMBER OF COMPLAINTS

The biggest divergence between the reports lies in the number of complaints from one company to another. From January to June 2018, including reports from both complaint bodies and individuals, Twitter counted 264,818 cases, YouTube 214,827 cases, and Facebook 886 cases filed explicitly as NetzDG complaints. For the second half of 2018 and for the same type of

complaints, Twitter counted 256,462 cases, YouTube 250,957 and Facebook 500. The gap between the figures published by Facebook and those published by Twitter and YouTube is still noticeable and shows no real change from the first round of reports. For the following analysis, the first round of reports can, therefore, be transferred to the period between July and December 2018.

These figures leave a big question mark hanging over the volume gap between YouTube and Twitter on the one hand, and Facebook on the other, especially since Facebook has the most users and the least complaints (absolutely and relatively). As already mentioned, the implementation of a complaint tool for users is crucial to the numbers that later constitute the central part of the report. After reading the reports and exploring the complaint mechanisms, the correlation between the flagging tool and the complaints filed seems obvious but should be analysed carefully. Compliance with section 3 (1) NetzDG bears the only meaningful difference between the social networks, hence its implementation is why Facebook's numbers are significantly lower. Regarding the implementation of a NetzDG complaint tool, two approaches can be observed: either the NetzDG complaint procedure is incorporated within the flagging tool of the social network or it is located somewhere else. In the latter approach, the usual flagging tool does not include the option "complaint under NetzDG" at first glance. While Google and Twitter chose to include the NetzDG complaint in their flagging tool (visible in the first step described above), Facebook placed the access to its NetzDG complaint procedure separately from the content under its imprint and legal information.

To be more specific, Facebook's complaint form according to NetzDG is not incorporated in their feedback function next to the contentious post. Users who want to report third-party content will first be offered the report categories under Facebook's community guidelines, when clicking on a button to "give feedback". Categories include nudity, violence, harassment, suicide or self-harm, fake news, spam, hate speech and illegal sales. In addition to Facebook's reporting tool, a complaint under NetzDG can be submitted by using an external link located next to Facebook's *"Impressum"* (the company's imprint and legal information). This begs the question of whether or not this type of implementation of sec. 3 NetzDG is sufficient, which I examin below. Before analysing the consequences of this implementation in the next section, one has to bear in mind that the NetzDG does not oblige social networks to incorporate their respective complaint procedure into pre-existing complaint mechanisms.

4. INHERENT BIAS DUE TO A DIVERGENT IMPLEMENTATION

According to sec. 3 (1) NetzDG, social networks have to implement an "easily recognisable, directly accessible and permanently available procedure for submitting complaints". As described above, this provision has been implemented in quite disparate manner by YouTube and Twitter, on the one hand, and Facebook, on the other, even though the complaint procedure might actually constitute the linchpin of the legislative project. As only a similar (if not identical) implementation of this provision would lead to comparable results, the data produced can barely be used as grounds for evaluation and development because of the disparate nature of its implementation. The small number of complaints through their NetzDG tool raises the question as to whether or not Facebook's complaint tool fulfils the requirements of sec. 3 (1) NetzDG and how it affects the significance of the reports.

4.1. USER-FRIENDLY COMPLAINT PROCEDURE

As aforementioned, to comply with sec. 3 (1) NetzDG, platforms do not have to connect existing reporting tools to the NetzDG procedure. The latter must, however, be user-friendly, that is, an "easily recognisable, directly accessible and permanently available procedure". The legislative memorandum does not provide further details on how these requirements shall be translated in the design of the complaint procedure. Hence, the question is not about the margin of discretion, but whether or not a procedure such as Facebook's is meeting these criteria. First of all, the link to the NetzDG complaint form is not easily recognisable for users since it is located far away from what users are able to immediately see when using the platform's complaints procedure. When a user sees a post that he or she believes to be unlawful, the feedback tool alongside it only shows the categories of the community guidelines and does not mention the possibility of reporting it under NetzDG provisions. The detour via an external link located next to Facebook's *"Impressum"* can hardly be described as "easily recognisable". Providing a NetzDG link next to a website's imprint is easily recognisable if you are looking for Facebook's general company information, but not if your goal is to report hate speech. That being said, it is "permanently available" when a user accesses Facebook in Germany.

Looking at the low volume of complaints under NetzDG in Facebook's case, one cannot help but connect the remote location of the complaint link and the numbers in Facebook's report. 886 cases in the first half-year of 2018 do not correlate with the high number of users and Facebook's constant struggle with unwanted content when it comes to hate speech (and other contentious posts) (Pollard, 2018). Libel, defamation or incitement to violence have been a constant issue for the world's largest social network (Citron & Norton, 2011, p. 1440) and Facebook only recently started to uncover some of its takedown rules (Constine, 2018). The latter have been kept secret for a long time, opening up speculation as to Facebook's real takedown policies. Social media platforms are often criticised for their lack of transparency when it comes to policing speech (Citron & Norton, 2011, p. 1441; Ranking Digital Rights, 2018 Index, Section 6).

Every Facebook user has access to a reporting tool (the feedback button next to a post), regardless of the NetzDG provisions, but he or she might not be aware of the additional possibility provided by the NetzDG – which makes this case quite special. Not only is this additional complaint procedure well hidden, but once a user is presented with the NetzDG complaint form (on Facebook), he or she will be warned that any false statement could be punishable by law (even if this rule doesn't apply to statements made to private parties). On the one hand, this might discourage people who wish to complain for no genuine reason and reduce the costs of unnecessary review loops. On the other hand, it could prevent users from reporting potentially unlawful content, which is cause for concern as it may result in chilling effects. The notion of chilling effects comes from US First Amendment scholarship and was introduced by a US Supreme Court ruling in 1952. The "chilling effects" concept essentially means that an individual will be deterred from a specific action under the "potential application of any civil sanction" (Schauer, 1978, p. 689). If a user – who is probably unsure of the lawfulness of third-party content – tries to use the complaint procedure and is confronted with the warning that any false statement could lead to legal steps, the act of deterrence seems likely. Again, Facebook's NetzDG report is too short and unspecified to infer from the simple numbers on chilling effects. The latter is nevertheless not to be underestimated and all these elements combined suggest that Facebook is pushing its users away from the NetzDG complaint procedure.

4.2. BYPASSING THE LEGISLATIVE GOAL?

The concrete implementation of this complaint procedure is decisive for its usability, but it is

also relevant for the informational value of the figures featured in the reports. This begs the question as to whether implementing the complaint procedure the way Facebook did could mean bypassing the legislative goal of the NetzDG. The implementation of the complaint procedure should in the first place – as stated above – protect users, not serve reporting purposes. A company's compliance with this obligation is therefore related only collaterally to the information value of the reports. Nonetheless, it is important when it comes to evaluating how the law contributes to the enhanced protection of users. The arguments above have shown that although social networks have fulfilled their transparency obligation by publishing reports, there is actually very little that we can conclude from them. The Facebook case makes it even more difficult to use the figures as the foundation for further development. Compliance alone does not lead to insightful data. Speculating on the reasons why Facebook decided to keep users away from using the NetzDG complaint procedure will not lead anywhere, but what is certain is that the law was implemented in a rather symbolic way. One may raise doubts on how accurate the formulation of sec. 3 (1) NetzDG regarding the implementation of an "easily recognisable" complaint procedure is. However, it would be too easy to blame it on the wording alone.

The argument was made that Germany could not expect platforms to "self-regulate in its [in Germany's] interest" (Fagan, 2017, p. 435) because of the relatively small size of its market on a global scale. On the contrary, if a company wants to do business in several countries it needs to respect each country's laws, especially if the laws in a specific country are, in principle, in line with the company's own guidelines – just as most of the offences enumerated in the NetzDG overlap with the "hate speech" category of platforms' community guidelines. In the case of Germany, there is no legal obligation to prioritise the relevant legal norms over the platform's community guidelines (as long as the unlawful content will be taken down), but that does not set aside the obligation to implement a user-friendly NetzDG complaint procedure. The subsequent question is: does Facebook's failure to implement an easily-recognisable complaint procedure according to sec. 3 (1) NetzDG mean that it is also bypassing the general legislative goal?

The answer is that, even though Facebook's complaint procedure is very likely to violate the legal provision (see supra, section 4.1), it has – in sum – achieved the outlined objective, that is, to remove hate speech quicker. The German government's overall goal in 2017 was to force social networks to be speedier in their responses to alleged offences on their platforms. That is why the time span for takedown decisions is limited to 24 hours and why any breach of the obligation to ensure that this type of fast-track procedure would be severely fined. The legislator wanted to remove unlawful content from sight as quickly as possible while ensuring the users' right to a due process. These reports confirm there was a need to address the issue of verbal coarsening and to protect digital communication spaces from hate speech. All three examined platforms name hate speech as the first source of complaints under the NetzDG. The social networks all implemented additional reporting tools (as part of the mandatory procedure), they deployed additional resources, and responded to the majority of complaints within 24 hours. To that extent the main requirements were met. The entry into force of the NetzDG led to larger and more specialised reviewer teams who could potentially provide a more granular review procedure. Under these circumstances it would be wrong to conclude that any of the platforms explicitly bypassed the primary legislative goal. Facebook's implementation nevertheless undermines the significance of the reports since the numbers produced cannot be taken into account for an advanced evaluation.

CONCLUSION

The discussion around content moderation by social media platforms and its regulation is still unresolved on many levels. More work needs to be done on the relation between private rules for content moderation and national laws, including the question of prioritisation. This is also true for the enforcement of rules and the role of non-human content review in that process. We have seen that platforms cannot solely rely on technology, such as upload filters, for example, to carry out the task of content moderation since the technology is still not fully capable of recognising hate speech. Although most of the criticism around the NetzDG with regards to constitutional law still remains valid, the reports analysed in this paper show that there are other aspects on which attention should lie, such as the implementation of complaint procedures. Unfortunately, the reports analysed constitute no reliable ground for screening and a further development of this obligation despite the data they contain. This is mainly due to the fact that the biggest player, Facebook, has dodged the obligation of creating an accessible and user-friendly NetzDG complaint procedure, preferring to manoeuvre users towards its own feedback form featuring its own categories of community standards. There was no change in this regard in the second round of reports. As long as platforms prioritise their own community rules, the effects on online speech remain more or less similar than before the coming into force of the NetzDG making it almost impossible to truly evaluate the impact of such regulation.

We can nonetheless wonder about the added value of an additional complaint tool within the platform's feedback mechanisms. Since (most) social media platforms operate globally, moderating content on the basis of global community guidelines is more cost-effective than if it was conducted on the basis of national regulation. Thus, the NetzDG reports could lead to the conclusion that this type of additional feedback tool, which would vary from country to country (because of national regulations), is ineffective and therefore unnecessary. As mentioned in the last section, the NetzDG did push the platforms to eventually take action against hate speech, an achievement which should not be downplayed. Perhaps, ensuring a faster review of user content by a more specialised content moderator is a sufficient goal for this type of law. The only conclusion to be drawn is that, for the time being and for the sake of acting on a global scale, social media platforms will prioritise their community guidelines when it comes to moderating user content.

I would like to thank Stephan Dreyer and Nikolas Guggenberger for their valuable feedback on the earlier draft of this paper. Thank you to the peer-reviewers for reading and evaluating this article.

REFERENCES

Balkin, J. M. (2018). Free Speech is a Triangle. *Columbia Law Review*, *118*(7), 2011-2056. Retrieved from https://www.jstor.org/stable/26524953

Balkin, J. M. (2014). Old-school/new-school speech regulation. *Harvard Law Review*, *127*(8), 2296-2342. Retrieved from https://harvardlawreview.org/2014/06/old-schoolnew-school-speech-regulation/

Belli, L., Francisco, P.A., & Zingales, N. (2017). Law of the Land or Law of the Platform? Beware of the Privatisation of Regulation and Police. In L. Belli & N. Zingales (Eds.), *Platform Regulations: How Platforms are Regulated and How They Regulate Us – Official Outcome of the UN IGF Dynamic Coalition on Platform Responsibility* (pp. 41-64). Rio de Janeiro: FGV Direito Rio Edition.

Buermeyer, U. (2017, March 24). Facebook-Justiz statt wirksamer Strafverfolgung?. *Legal Tribune Online*. Retrieved from https://www.lto.de/recht/hintergruende/h/netzwerkdurchsetzungsgesetz-netzdg-facebook-strafverfolgung-hate-speech-fake-news/

Celeste, E. (2018). Terms of service and bills of rights: new mechanisms of constitutionalisation in the social media environment? *International Review of Law, Computers & Technology*, *33*(2), 122-138. doi:10.1080/13600869.2018.1475898

Citron, D. K. (2017). Extremist Speech, Compelled Conformity, and Censorship Creep. *Notre Dame Law Rev.*, *93*(3), 1035-1071. Retrieved from https://scholarship.law.nd.edu/ndlr/vol93/iss3/3/

Citron, D. K., & Norton, H. L. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, *91*(4), 1435-1484. Available at http://www.bu.edu/law/journals-archive/bulr/volume91n4/documents/CITRONANDNORTON.pdf

Constine, J. (2018, April 24). Facebook reveals 25 pages of takedown rules for hate speech and more. *TechCrunch*. Retrieved from https://techcrunch.com/2018/04/24/facebook-content-rules/?guccounter=2

Delort, J. Y., Arunasalam, B., & Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, *15*(3), 9-30. doi:10.2753/JEC1086-4415150302

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. *Proceedings of the 24th International Conference on World Wide Web, 29–30*, 29-30. doi: 10.1145/2740908.2742760 Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.697.9571&rep=rep1&type=pdf

Fagan, F. (2017). Systemic Social Media Regulation. *Duke Law & Technology Review*, *16*(1), 393-439. Retrieved from https://scholarship.law.duke.edu/dltr/vol16/iss1/14/

Funke, D. (2018, July 24). A guide to anti-misinformation actions around the world, *Poynter*. Retrieved from https://www.poynter.org/news/guide-anti-misinformation-actions-around-

world

Gaydhani A., Doma V., Kendre S., & Bhagwat L. (2018). Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. *arXiv:1809.08651v1 [cs.CL]*. Retrieved from https://arxiv.org/pdf/1809.08651.pdf

Gersdorf, H. (2017). Hate Speech in sozialen Netzwerken – Verfassungswidrigkeit des NetzDG-Entwurfs und grundrechtliche Einordnung der Anbieter sozialer Netzwerke. *MMR – MultiMedia und Recht,* (7), 439-447.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gollatz, K., Riedl, M. J., & Pohlmann, J. (2018, August 9). Removals of online hate speech in numbers [Blog post]. Retrieved from HIIG Digital Society Blog: https://www.hiig.de/en/removals-of-online-hate-speech-numbers/ doi:10.5281/zenodo.1342325

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All You Need is "Love": Evading Hate-speech Detection. *arXiv preprint arXiv:1808.09115*. Retrieved from https://arxiv.org/pdf/1808.09115.pdf

Guggenberger, N. (2017a). Das Netzwerkdurchsetzungsgesetz – schön gedacht, schlecht gemacht [The Network Enforcement Act – well thought, poorly done]. *Zeitschrift für Rechtspolitik*, 2017(4), 98-101.

Guggenberger, N. (2017b). Das Netzwerkdurchsetzungsgesetz in der Anwendung [The Network Enforcement Act in application]. *Neue Juristische Wochenschrift*, 36, 2577-2582.

Holznagel, D. (2018). Overblocking durch User Generated Content (UGC)-Plattformen: Anspru¨che der Nutzer auf Wiederherstellung oder Schadensersatz? *Computer und Recht*, *34*(6) 369-378. doi:10.9785/cr-2018-340611

Kaye, D. (2018) Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, United Nations Human Rights Council, *A/HRC/38/35*. Retrieved from http://undocs.org/A/HRC/38/35

Keller, D. (2018). *Internet platforms: Observations on speech, danger, and money* (Aegis Series Paper No. 1807). Stanford, CA: Hoover Institution. Retrieved from: https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf

Klonick, K.(2018). The New Governors: The People, Rules, and Processes Governing Online Speech, *Harvard Law Review*, *131*(6), 1598-1670. Retrieved from https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Koebler, J., & Cox, J. (2018, August 23). The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. *Vice Motherboard*. Retrieved from https://motherboard.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

Liesching, M. (2018a). Lösungsmodell regulierter Selbstregulierung – Zur Übertragbarkeit der

JMStV-Regelungen auf das NetzDG. In M. Eifert, T. Gostomzyk (Eds.), *Netzwerkrecht*, (pp. 135 – 152). Baden-Baden: Nomos.

Liesching, M. (2018b). Die Durchsetzung von Verfassungs- und Europarecht gegen das NetzDG. *MMR – MultiMedia und Recht*, (1), 26-30.

Liesching, M. (2018c). Netzwerkdurchsetzungsgesetz 1. Online-Auflage. Baden-Baden: Nomos.

Matsakis, L. (2018, September 26). To Break a Hate-Speech Detection Algorithm, Try 'Love'. *WIRED*. Retrieved from https://www.wired.com/story/break-hate-speech-algorithm-try-love/

Newton, C. (2019, February 25). The Trauma Floor. The secret lives of Facebook moderators in America. *The Verge*. Retrieved from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Nolte, G. (2017). Hate-Speech, Fake-News, das »Netzwerkdurchsetzungsgesetz« und Vielfaltsicherung durch Suchmaschinen. *ZUM*, 7, 552-565.

Nunziato, D. C. (2014). The Beginning of the End of Internet Freedom (Law School Public Law Research Paper No. 2017-40). Washington DC: George Washington University. Available at https://scholarship.law.gwu.edu/faculty_publications/1280/

Pollard, A. (2018, July 05). Facebook Found “Hate Speech” in the Declaration of Independence, *Slate*. Retrieved from https://slate.com/technology/2018/07/facebook-found-hate-speech-in-the-declaration-of-independence.html

Ranking Digital Rights, (2018). *Corporate Accountability Index*. Retrieved from https://rankingdigitalrights.org/index2018/report/executive-summary/

Read, M. (2019). Who Pays for Silicon Valley’s Hidden Costs? *New York Magazine*. Retrieved from http://nymag.com/intelligencer/2019/02/the-shadow-workforce-of-facebooks-content-moderation.html

Reinhardt, J. (2018, January 15). A Slight Case of Overblocking: Les enjeux constitutionnels de la loi allemande sur les réseaux sociaux [The constitutional issues of the German social networks law] [Blog post]. Retrieved from Jus Politicum http://blog.juspoliticum.com/2018/01/15/a-slight-case-of-overblocking-les-enjeux-constitutionnels-de-la-loi-allemande-sur-les-reseaux-sociaux-par-jorn-reinhardt/

Reuter, M. (2018). Das Netzwerkdurchsetzungsgesetz gefährdet die Meinungsfreiheit [The Network Enforcement Act endangers freedom of opinion]. In T. Müller-Heidelberg, M. Pelzer, M. Heiming, C. Röhner, R. Gössner, M. Fahrner, H. Pollähne, & M. Seitz (Eds.) *Grundrechte-Report*. Frankfurt am Main: Fischer Taschenbuch Verlag.

Richter, P. (2017). Das NetzDG – Wunderwaffe gegen „Hate Speech“ und „Fake News“ oder ein neues Zensurmittel?, *ZD-Aktuell*, 9, 05623

Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo Law Review*, *24*(4), 1523-1467. Retrieved from https://larc.cardozo.yu.edu/faculty-articles/148/

Schauer, F. (1978). Fear, risk and the first amendment: Unraveling the chilling effect. *Boston University Law Review*, *58*, 685-732. Available at https://scholarship.law.wm.edu/facpubs/879/

Schiff, A. (2018). Meinungsfreiheit in mediatisierten digitalen Räumen - Das NetzDG auf dem Prüfstand des Verfassungsrechts, *MMR – MultiMedia und Recht,* (6), 366-371.

Schulz, W. (2018). *Regulating Intermediaries to Protect Privacy Online – the Case of the German NetzDG* (Discussion Paper No. 2018-01) Berlin: Alexander von Humboldt Institut für Internet und Gesellschaft. Retrieved from https://www.hiig.de/publication/regulating-intermediaries-to-protect-privacy-online-the-case-of-the-german-netzdg/

Schulz, W., & Held, T. (2002). *Regulierte Selbstregulierung als Form modernen Regierens. Im Auftrag des Bundesbeauftragten für Angelegenheiten der Kultur und der Medien. Endbericht* [Regulated self-regulation as a form of modern governing. On behalf of the Federal Commissioner for Cultural and Media Affairs. Final report] (Working Paper No. 10). Hamburg: Verlag Hans-Bredow-Institut. Retrieved from https://www.hans-bredow-institut.de/uploads/media/Publikationen/cms/media/a80e5e6dbc2427639ca0f437fe76d3c4c95634ac.pdf

Scott, C. (2004). Regulation in the age of governance: The rise of the post-regulatory state. In J. Jordana, & D. Levi-Faur (Eds.), *The politics of regulation: Institutions and regulatory reforms for the age of governance* (pp. 145-173). Cheltenham: Edward Elgar. doi:https://doi.org/10.4337/9781845420673.00016

Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword? Retrieved from https://ssrn.com/abstract=3306964

Wimmers, J., & Heymann, B. (2017). Zum Referentenentwurf eines Netzwerkdurchsetzungsgesetzes (NetzDG) - eine kritische Stellungnahme. *AfP - Zeitschrift für das gesamte Medienrecht*, *48*(2), 93-102. doi:10.9785/afp-2017-0202

Wischmeyer, T. (2018). 'What is Illegal Offline is Also Illegal Online' – The German Network Enforcement Act 2017. doi:10.2139/ssrn.3256498

YouTube (2017, December 4). Expanding our work against abuse of our platform. Retrieved from https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

FOOTNOTES

1. The NetzDG itself might not be flawless but in some way, it can serve as an experiment for other lawmakers.

2. NetzDG quotes are from the official translation by the German Ministry of Justice.

3. Loose translation of the NetzDG explanatory memorandum, retrieved from https://dipbt.bundestag.de/doc/btd/18/123/1812356.pdf, last accessed 13 May 2019.

4. This paper does not examine the case of the social network Google+ explicitly, but its report shows that the legal requirements according to NetzDG were implemented in the same manner as for YouTube. The platform Change.org is also within the scope of application but was not examined here for the purpose of focusing on the biggest platforms.