Social media governance and strategies to combat online hatespeech in Germany

Received: 15 December 2022 | Accepted: 17 May 2023

DOI: 10.1002/poi3.348

RESEARCH ARTICLE

P&I Policy & Internet | PSO | WILEY

Social media governance and strategies to combat online hatespeech in Germany

Daniela Stockmann[1] | **Sophia Schlosser**[2] | **Paxia Ksatryo**[1]

[1]Centre for Digital Governance, Hertie School, Friedrichstrasse, Germany

[2]ETH Zürich, Zürich, Switzerland

Correspondence

Daniela Stockmann, Centre for Digital Governance, Hertie School, Friedrichstrasse 180, 10177 Berlin, Germany.
Email: stockmann@hertie-school.org

Funding information

German Ministry for Family, Senior Citizens, Women, and Youth (Bundesministerium für Familie, Senioren, Frauen und Jugend, BMFSFJ)

Abstract

Concerns over online hatespeech have prompted governments to strengthen social media governance. However, claims by policy-makers and political activists regarding the effectiveness and likely consequences of legal regulations remain largely untested. We rely on qualitative interviews and two expert surveys to examine the behavior of public relations professionals in response to online hatespeech when having the option of using the new user-complaint mechanism under the German Network Enforcement Act (NetzDG). Our findings reveal that strategies depend on whether professionals work at public sector institutions, business, or civil society organizations and political parties. Public sector institutions are likely to report to the platform, but not under NetzDG. Civil society organizations are likely to choose content moderation, counterspeech, and other forms of intervention. Businesses deploy a wide range of strategies. In practice, Germany's procedural approach relying on user-complaint mechanisms to deal with online hatespeech is not used by experts as a means to combat online harassment.

KEYWORDS

Germany, governance, hatespeech, public relations, social media

INTRODUCTION

Online hatespeech is a pressing issue on social media. Concerns over civil public discourse and the spread of online harassment have prompted governments to take action in New Zealand, the United Kingdom, the United States, and India, to name just a few examples (Bayer et al., 2021; Persily & Tucker, 2020), and many policy-makers are turning toward coregulatory approaches (Vermeulen, 2019). Coregulation describes a governmental strategy that entails more state involvement than self-regulation, whereby the state sets up an explicitly nonstate regulatory regime that includes technology companies, but also advertisers and consumer organizations or other public interest groups (Kleinsteuber, 2004). Advantages of this form of regulation include offloading costs for monitoring and setting standards to industry. Simultaneously, industry is incentivized to comply with public interest in sight of stronger accountability and compliance mechanisms imposed by the state (Lodge & Wegrich, 2012).

In Europe, Germany has been at the forefront of these developments, with France and Austria later passing similar laws (Bayer, 2022; Vermeulen, 2019). More recently, the European Union has passed the Digital Services Act/Digital Markets Act package under the leadership of German President of the European Commission Ursula von der Leyen. These new approaches toward coregulation strengthen the role of public institutions by increasing oversight over social media while also strengthening users' procedural rights as citizens, which encompass the right to court, fair trial, and an effective remedy (Bayer, 2022). Procedural safeguards are written into the law to assure that the state, not private companies, has the final verdict. Observers have therefore described the German and European laws as a process-based approach embodying "new school speech regulation" (Balkin, 2018; Griffin, 2022; Klausa, 2023).

One important procedural safeguard constitutes a user-complaint mechanism, which was first written into the German Network Enforcement Act (NetzDG), passed in June 2017. NetzDG obliges social networks with above two million users in Germany "to supply users with an easily recognizable, directly accessible, and permanently available procedure for submitting complaints about unlawful content" (Bundesministerium der Justiz, 2017). Once notified by a user or complaint body, the platform must remove "manifestly unlawful content" within 24 h (with some discretion) or they risk facing fines of up to 50 million euros. Platforms that receive over 100 complaints per calendar year must publish biannual transparency reports. When implementing NetzDG, Facebook, Twitter, and YouTube provided users with a German IP address with a new option to report illegal content under NetzDG; this option is separate from reporting violations under the platform's own community standards. Upon nonresponse by the platform, users are also given the right to report to the German Ministry of Justice with the aim of empowering the state to place pressure on social media to strengthen moderation of illegal content. Given these parameters, the successful implementation of NetzDG can be said to rely on user submissions.

In this article, we examine the extent to which key societal actors who face hatespeech online actively invoke procedural rights empowered by a new coregulatory framework for social media governance. We concentrate on Germany because we have the opportunity to examine the initial implementation phase, starting in 2018, and after revisions are made to improve user-friendliness in 2021 (Bundesministerium der Justiz, 2021). This allows us to investigate a longer time span. Based on 18 qualitative interviews and two rounds of expert surveys with public relations (PR) professionals, we provide needed insights into the effectiveness of user-complaint mechanisms based on empirical evidence. In the case of NetzDG, German policy-makers initially framed the law as an attempt to address increasing online hatespeech (Bundesministerium der Justiz, 2017). Building on research in public

policy that evaluates policy implementation based on the extent to which targets actively utilize policies (Patton, 2008, 2012), we concentrate on PR professionals who specialize in social media management. Being responsible for an organization's online presence and communications strategy, they are experts on content moderation and strategies to deal with online hatespeech while also representing organizations as important stakeholders in policy implementation.

We argue that successful implementation of NetzDG with user-complaint mechanisms at its core rest on the logic of coproduction defined by Ostrom (1996) as "the process through which inputs used to produce a good or service are contributed by individuals who are not 'in' the same organization." These individuals (often referred to "citizens, clients, consumers, volunteers, and community organizations") simultaneously produce the public service while also benefiting from them (Alford, 1998). Our study reveals that only a small minority of the targeted societal actors—in this case, PR professionals—actively make use of user-complaint mechanisms. Only those working for public sector institutions tend to report to platforms, but without invoking NetzDG. Civil society organizations and political parties prefer content moderation and counterspeech while businesses deployed a wide range of strategies. These findings have implications beyond the scholarly discussion on NetzDG (Eifert et al., 2020; Heldt, 2019; Liesching, 2020; Tworek & Leerssen, 2019) and contribute to scholarly and policy discussions on the effectiveness of the new European approach toward social media governance. More broadly, our findings advance existing discussions on content moderation and coproduction to include the professional world.

LITERATURE REVIEW

User strategies to combat online hatespeech

Although hatespeech is often discussed in public and academic discourse in relation to freedom of speech and citizen responsibilities (Howard, 2019), the concept often remains opaque (Sellars, 2016). Parekh (2012, p. 41) characterizes the following key elements: first, hatespeech "is directed against a specified or easily identifiable individual, or more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature"; second, it stigmatizes the victim by "ascribing to it qualities widely regarded as undesirable"; and finally, it projects the victim as "an undesirable presence and a legitimate object of hostility." *Intention* to promote hatred, violence, or resentment must be evident, which in turn leads to *consequences* (Massey, 1992; Parekh, 2012). Aside from the importance of interpersonal contexts in defining hatespeech, scholars have emphasized that intention of hatespeech content is difficult to prove or demonstrate empirically. Hatespeech content is "the use of certain symbols and iconography […] or words or phrases generally considered to be offensive to a particular minority" (Marwick & Miller, 2014). In light of this literature, we define hatespeech as an action that intentionally violates a person or a group with communication. It is distinct from other communication by its degree of severity, which is indicated by its specific context, by its publicity, and by its consequences. Our qualitative interviews, explained in more detail below, confirmed that this conceptualization matches how PR professionals in Germany identify hatespeech online.

While much of the discussion on online hatespeech focuses on defining and detecting its presence, a growing body of scholarship identifies strategies to combat hatespeech and online harassment (Davies et al., 2016; Ganesh & Bright, 2020; Szmania & Fincher, 2017). The most common strategy to combat online hatespeech is *content moderation*. Content

moderation entails different types of actions, for example, removing content, deactivating comments, or blocking users based on the community guidelines defined by social media platforms (Ganesh & Bright, 2020; Gillespie, 2018; Roberts, 2019). As the rate of content uploaded exponentially outpaces the speed at which human moderators can review content, online content providers have co-opted *algorithmic moderation* (Gorwa et al., 2020). Scholars have identified a number of challenges associated with content moderation: first, since no common definition of hatespeech exists, content moderation still lacks consistency within and across platforms (Douek, 2021). Users who are censored on one platform locate online spaces where they are not, as illustrated by the move from Trump supporters away from Twitter toward Parler. Language is also context-dependent and it is complex to correctly identify hatespeech via algorithmic and human moderation (Gorwa et al., 2020). In addition, numerous observers have pointed out platforms applying community standards inconsistently. For example, during the coronavirus disease (Covid) 2019 pandemic, Facebook decided to ban ads for medical masks, but the bans did not apply to all commercial ads and instead blocked instructions on how to sew cloth masks at home (Douek, 2021). Other critics point to the slow response of platforms in moderating harmful content, leading to self-organized efforts on the side of minorities to take action into their own hands (Díaz & Hecht-Felella, 2021). Overall, due to inconsistencies and slow response times, content moderation often suppresses "valuable as well as illicit speech" (Hangartner et al., 2021, p. 1).

A main alternative to content moderation constitutes intervention by third parties. The most widely discussed form of intervention is *counterspeech*, which aims to reduce or avoid hate through interaction, communication, and persuasion of the author of hateful content (Douek, 2021). For example, Hangartner et al. (2021) focus on the three most common strategies for counterspeech against xenophobic or racist content: using humor, inducing empathy for others, and alerting others to the consequences of hatespeech. Evidence based on a field experiment on Twitter demonstrates small but consistent effects for preventing or reducing hatespeech. Especially, empathy "seeks to humanize the victim and remind the sender that people can be hurt by her behavior" (Hangartner et al., 2021, p. 1). Other forms of intervention are also considered as part of the research on cyberbullying (Beutel et al., 2016; Van Eerten & Doosje, 2019). Bystander intervention can reduce the amount of cyberbullying victimization on social media, but bystanders often fail to act (DiFranzo et al., 2018; Taylor et al., 2019). Group processes are often crucial, and the visibility and acceptance by outsiders play a highly significant role beyond the bully and victim (Salmivalli et al., 1996). Studies demonstrate that providing information about the size of the audience and further nudges mobilize intervention by social media users (DiFranzo et al., 2018; Taylor et al., 2019). These constitute some mechanisms that platforms may offer to support intervention on social media.

We advance this body of research in Political Communication in two ways: first, we expand current research to include professional users working for organizations. Second, we add reporting under NetzDG as a new reporting mechanism to the discussion and examine its consequences on the effectiveness of combating online hatespeech (see Figure 1). Investigating the use of this new reporting mechanism provides insights into its potential effectiveness. Use constitutes a precondition for complaint mechanisms to have an effect on how platforms manage online hatespeech. To our knowledge, prior research has not examined the use of NetzDG reporting mechanisms—in part due to incomplete information on transparency reports (e.g., Heldt, 2019):

> To what extent are reporting options provided under NetzDG used by PR professionals in an effort to address online hatespeech?

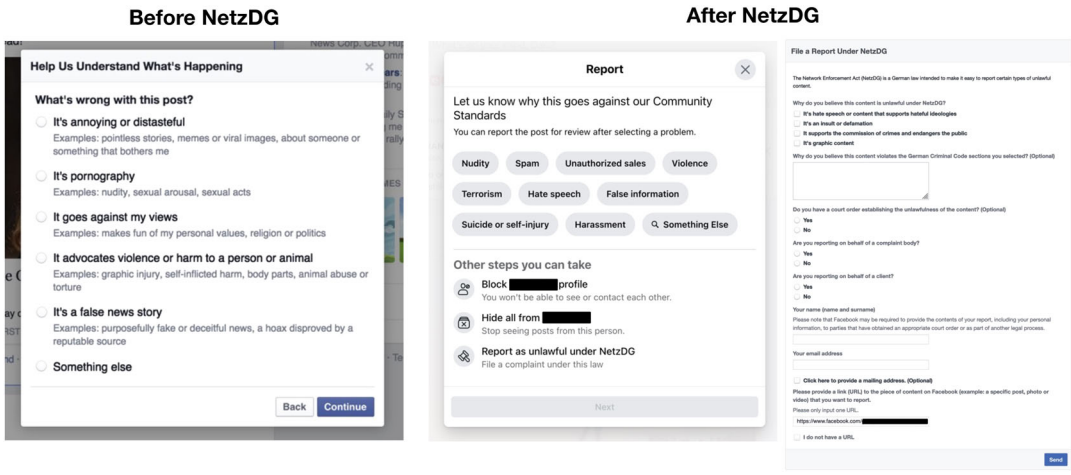

FIGURE 1 Reporting option on Facebook pre- and post-German Network Enforcement Act (NetzDG). *Source*: Screenshot by the Authors.

Role of organizations in coproduction

Organizations and their respective interests are core to policymaking. Olson's (1971) classic definition rests on the assumption that an organization's main purpose is to advance the most common interests of a group of individuals. The achievement of a group's common interest implies that a public or collective good has been provided to that group (Olson, 1971). The common interest advanced by organizations is representative of various societal dimensions (cultural, political or social issues, ethnic or religious communities, economic interests), with socioeconomic organizations most represented in policymaking (Hassel, 2010). Organizations can also exist in hybrid forms, such as social enterprises that want to address social problems and achieve corporate goals simultaneously (Mair & Seelos, 2021).

These types of organizations, including government agencies, societal organizations, and corporations, affect each stage of the policy process and assume different roles at different stages (Kraft & Furlong, 2010; Sutton, 1999). For example, Vaughan and Arsneault (2013) argue that nonprofits play four roles in the policy-making process: first, they make policy by—for instance—direct service delivery; second, they influence policy by bringing attention to public problems and proposing solutions; third, they are the subject to policy requirements when they implement policy in contract with the government; finally, they are affected by the policy when evaluating how the government has addressed a public problem. More specifically, organized interest groups and corporations lobby in the agenda-setting stage, provide expertise or submit positions during policy formulation and deliberation stages via public consultation processes, testify in legislative hearings, act as implementing agencies for a policy, and may be the subject of regulation (Hassel, 2010; Kraft & Furlong, 2013).

We examine a case in which organizations from different sectors play a role in coproduction. While coproduction originally emphasized the role of the citizen (Warren et al., 1984), its meaning has grown to also include quasiprofessionals and organizations, mainly community groups, associations, and nonprofits (Alford, 1998; Howlett, 2019; Poocharoen & Ting, 2015). In the case of NetzDG, all organizations who maintain a social media presence can also be classified as a coproduction agent. Most of the takedowns on

platforms are de facto detected by automated content moderation tools (for more, see Bloch-Wehba, 2020; Gorwa et al., 2020). In addition, most social networks complement automated content moderation tools with voluntary monitoring by users to proactively take down content and users that do not comply with community standards. Since these mechanisms alone have been regarded as inefficient in addressing, for example, online hatespeech, NetzDG aims to increase pressure on social media platforms to moderate illegal online content by providing additional reporting options to users as a general category. Although the letter of the law was drafted without specific references to the end user, PR professionals are included in this category. PR professionals hold a duality: as online communication experts they are expert social media users; at the same time, they also represent organizations as important stakeholders in policy implementation. This duality places them at the core of coproduction.

By investigating the role of PR practitioners at the intersection between social media users and organizations advances existing research on the role of technology in coproduction, which has so far concentrated on understanding how the use of technology can aid or inhibit coproduction with citizens offline (e.g., Huang & Yu, 2019). For example, Jones et al. (2016) discuss a government-issued app "MapLocal" that crowd-sources local knowledge to facilitate neighborhood planning. Here we investigate how a user-complaint mechanism implemented via technology impacts coproduction *online* by potentially changing organizational strategies to manage communication on social media. Instead of focusing on users as a general category, we draw attention toward systematic differences that are expected among professional users. Existing research on differences between the public, private, and nonprofit sectors in coproduction leads us to expect systematic differences across sectors.

Public sector entities in the coproduction literature range from individuals such as healthcare workers and teachers to organizations at the community and government levels. Public sector professionals who coproduce are identified as persons possessing "specific knowledge and expertise […] a certain degree of autonomy" and work in tandem with citizens (Steen & Tuurnas, 2018, p. 80). Scholars have claimed that the success of coproduction, when including these servants, can also depend on the relationship between staff and users (Ostrom, 1996; Steen & Tuurnas, 2018). We may be inclined to believe that the public sector acts in the public interest, depending on how public professionals position their responsibilities in relation to citizens (Thomas et al., 1999). Alford (2002, p. 37) argues that public sector personnel tend to be motivated by willingness (intrinsic rewards, sociality, and expressive values) and ability, with less of an emphasis on material rewards sanctions.

Meanwhile, experts from civil society organizations can be viewed as natural coproducers given their community ties to citizens and position to represent public interest (van Eijk & Gascó, 2018). Because of the opportunity "to increase the amount and/or quality of service they enjoy" (Brudney & England, 1983, p. 60), civil society organizations tend to be oriented toward harmonizing their relationship with their respective target audience among citizens.

As the private sector is not formally explored in coproduction literature, we take inspiration from the discussion on the devolution of public services away from the state to the private sector (privatization) to explain why we examine behavioral differences among sectors. Brudney (1987) argues that businesses are similar to civil society organizations in their desire to shift the balance of meeting community needs away from state control, thus increasing its autonomy. However, the private sector diverges into two main points: first, the private sector tends to be committed to competition and production efficiency, unlike the other two sectors. Second, they differ in defining who their community is: while the private sector may gain a sense of public responsibility (Perry & Rainey, 1988), businesses are mainly influenced by their investors and profit-making self-interest (Kuczynski, 1999).

As a result of these structural differences, we investigate if sectoral underpinnings play a role in coproduction of NetzDG:

> What is the relationship between an organization's sector and the choice of counterstrategies among PR professionals?

RESEARCH METHODS

Research design and sampling

We conducted two online expert surveys among PR professionals in charge of managing social media, henceforth referred to as "PR professionals." The first was conducted after NetzDG's initial implementation in 2019 and the second after NetzDG's amendments were implemented in 2022. Comparing the 2019 and 2022 results provides us with an opportunity to test whether the use of NetzDG reporting options has increased over time. Expert surveys like ours are not intended as representative samples of the survey population, but instead, attempt to be inclusive of the universe of experts and possible factors that may influence their judgments—in our case working for different organizational sectors (Maestas, 2018).[1] To develop and design valid survey measures, we conducted 18 semistructured qualitative interviews with PR professionals before survey respondents were contacted via snowball sampling. Recruitment was supported by the German Federation of All PR professionals (*Bundesverband Deutscher Pressesprecher*) based on their list server, which includes the vast majority of PR professionals in Germany. To reach those PR professionals managing social media, we framed the survey as being aimed at hatespeech on social media; survey responses confirm that respondents felt knowledgeable about their organizations' social media strategies.[2] Furthermore, descriptive statistics confirm that each expert survey reached experts across categories that likely matter for their judgment: each survey includes significant percentages of experts working in the three sectors: almost half worked in businesses (2019: 49.82%; 2022: 41.12%), followed by public sector institutions (2019: 26.15%; 2022: 39.12%), and civil society organizations and political parties (2019: 24.03%; 2022: 19.76%). Similarly, experts ere located in all 16 states participated in both surveys. In both surveys, Facebook was used the most (2019: 77.74%; 2022: 77.23%), followed by Twitter (2019: 65.37%; 2022: 64.33%), Instagram (2019: 48.06%; 70.97%), and YouTube (2019: 66.25%; 2022: 68.50%), as well as LinkedIn (2019: 41.87; 2022: 61.10%).

Measurement

To facilitate the interpretation of results, all variables were coded to run from zero to one. We use listwise deletion on missing values. Both surveys are consistent in measurement.

Prior experience with hatespeech

Surveys asked whether respondents had experienced hatespeech at work. Based on scholarly literature and qualitative interviews, hatespeech was defined as "a communicative act aimed at degrading and defaming a person or group. […] The mere use of certain words is not sufficient to classify a statement as hatespeech ('I hate people who cannot listen'). The severity of the statement plays an important role, which is associated with the specific

context, the audience, and its consequences, ranging from weak cases (e.g., 'migrants are lazy!') […] to extreme cases ('He should drown together with the refugees!')."

Counterstrategies to hatespeech

Those who reported having experienced hatespeech were asked how they responded after the most recent incident. Based on existing literature and qualitative interviews, we developed a list of 13 strategies, which we found to be comprehensive across usage of different platforms. According to our interviews, all 13 strategies were available on Twitter, Instagram, Facebook, YouTube, and LinkedIn. *Reporting under NetzDG* is a dummy variable coded one for respondents choosing the option "report to platform as violation of NetzDG" or "report to federal office of justice as violation of NetzDG." *Prior reporting mechanisms* is a dummy variable for reporting options existing before NetzDG, including "report to platform as part of breach of community standards" and "file a criminal complaint." Content moderation options included "deleting comments," "hiding comments," and "deactivate comment function." Options for intervention and counterspeech included "refer users to the platform's etiquette," "refer users to own guidelines," "answer in an informative manner," "answer with counter-speech," and "mobilize own supporters." Finally, nonaction was listed as "ignore comments."

Sector

Dummy variables for the organization the respondent is working at include *Business, Public sector institution*, and *Civil society organizations and political parties*. These categories are regularly used by the German Federation of All PR professionals. Due to the small number of responses from press officers working for political parties ($N = 5$, which corresponds to the number of the largest parties in Germany), there are not enough cases to draw statistical conclusions about political parties as a separate category. Thus, civil society organizations and political parties were merged into one category. Since both types of organizations share similar structural motivations that bind them toward certain subgroups of citizens, such a categorization matches the rationale for investigating organizations by sector outlined above. Robustness checks reveal that strategies by civil society organizations and political parties indeed share similarities in terms of their reporting behavior.

Control variables

Engagement with audience captures the level of interactive engagement with the audience based on the question: "What is the goal of communication in the channel you use the most?"—one-way communication whereby interaction in terms of likes and shares are welcome; comments and questions of users are welcome and are responded to indicating somewhat interactive engagement; or active engagement with users to discuss content (multiple choice). On average, PR strategies were somewhat interactive (mean 2019 = 0.45; 2022 = 0.36). *Work related to political and societal debate* was based on a self-reported item with higher values indicating a stronger relationship (mean 2019 = 0.69, 2022 = 0.63). *Size of organization* was assessed based on self-reports by the respondent, running from small (0–9 employees) to large (more than 10,000 employees). The distribution of organization size for both surveys varied, on average, between 250 and 999 employees (mean 2019 = 0.51, 2022 = 0.49). In addition, we control for being *Based in East Germany*

(neue Bundesländer, mean 2019 = 0.07, 2022 = 0.09) and *Regional trend in hatespeech*, which captures the average subjective assessment of changes in hatespeech by respondent within the same federal state in the survey. Higher values indicate that respondents in the same federal state perceived hatespeech, on average, to have grown over the past 10 years. On average, respondents perceived hatespeech to be increasing (mean 2019 = 0.77, 2022 = 0.84). Finally, we control for *Female*, *Age*, and *Consequences on mental health*, which was based on a 5-step Likert scale on a question about how much the respondents and her colleagues had been mentally burdened by the last incident of hatespeech, with higher values indicating stronger burden (mean 2019 = 0.33, 2022 = 0.39).

RESULTS

Prevalence of online hatespeech

In qualitative interviews, PR professionals working in all sectors gave examples of cases they encountered at work. One expert was confronted with hatespeech when a company owner decided to train refugees as an integration effort. This was actively promoted, but people wrongly assumed that available training spots for Germans were reduced, triggering racist comments. These involved people from the right-wing political spectrum, but also employees of the company itself. The company reacted by deleting hateful comments and blocking users. Internally, the company owner spoke at all company locations, shaming haters and threatening to fire everyone conducting hatespeech (Interview #6, 2018). Another incident occurred at a gaming company that was involved in the production of "shooter games"; after a school shooting, the company started to receive death threats. In this case, the company decided to ignore messages to not increase attention (Interview #10, 2018). The announcement of a prize by a civil service organization triggered harassment and death threats when the prize was given to the captain of an organization that rescued refugees on boats on the Mediterranean Sea. The expert felt initially overwhelmed by the high quantity of severely negative and hateful comments. The organization decided to hide or delete the most extreme comments and to block users and report them to Facebook. They also repeatedly answered and reminded users to keep Facebook's and their own community standards in mind when interacting on Facebook and on their own homepage (Interview #11, 2019). In a public administrative body of a major German city, the PR officer was surprised that even seemingly nonsensitive content, such as references to solidarity, triggered criticism of openness toward hosting refugees in the city on social media. The organization chose to predominantly report inappropriate comments to Twitter, but reporting was reduced to only the most severe cases after a lack of reaction by the platform (Interview #10, 2019).

Survey results indicate that these incidents have become a serious matter of concern to PR professionals more generally. Around 45% of respondents reported having been confronted with online hatespeech in 2019 and 51% in 2022. Most respondents were confronted with hatespeech cases at least once within the past 3 months or earlier during the past year (2019: 63.64%; 2022: 61.97%); roughly 20% reported experiencing online hatespeech at least 1 per month (2019: 20.5%; 2022: 21.13%). In 2019, 16% stated to have been confronted with hatespeech on a weekly basis, 21% in 2022. Experts were particularly likely to encounter online hatespeech at public sector institutions (2019: 59.46%; 2022: 58.76%) as well as civil service organizations and political parties (2019: 55.88%; 2022: 60.20%); at businesses a smaller but still significant percentage had faced online hatespeech at work (2019: 31.92%; 2022: 40.19%). More generally, the vast majority perceived hatespeech to have increased over the past 10 years (2019: 64.59%; 2022:

71.50%), with most people perceiving the increasing trend to start in 2015 (2019 survey) and 2020 (2022 survey).

Use of reporting under NetzDG and alternative counterstrategies

To what extent are reporting options provided under NetzDG used by PR professionals in an effort to address online hatespeech? Figure 2 provides an overview of the different strategies to combat online hatespeech among PR professionals. Sixteen months after NetzDG had been enacted, the most commonly employed strategy to counter hatespeech was to answer informatively, while the reporting to the Federal Office of Justice as a violation of NetzDG was the least employed option. As of 2019, PR professionals were unlikely to use the new reporting mechanisms provided under NetzDG. Only 6% of experts relied on NetzDG reporting options, and this percentage only slightly increased over time.

Despite amendments to NetzDG in June 2021, which demanded platforms improve the user-friendliness of user-complaint mechanisms, only 11% of PR professionals chose NetzDG reporting options. In addition to reporting, two-sample *t*-tests with unequal variance

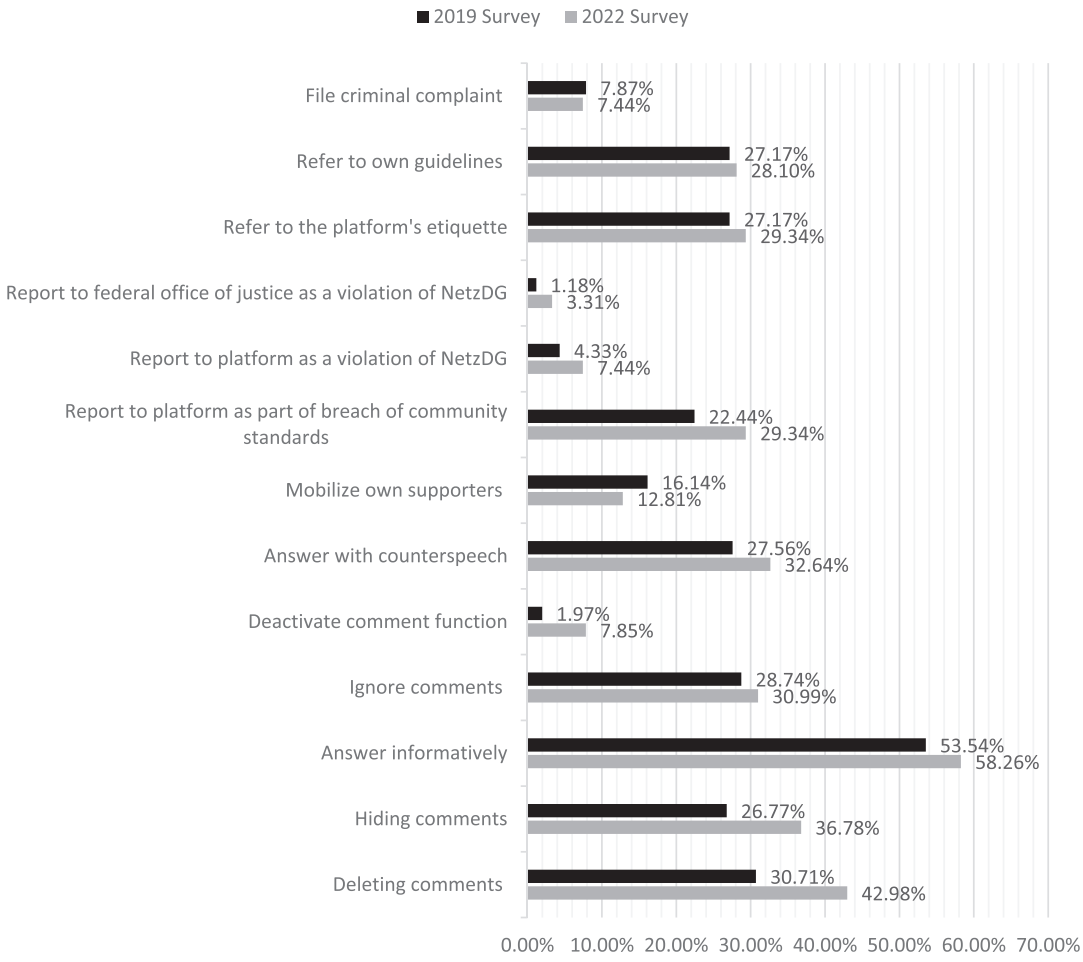

FIGURE 2 Counterstrategies to online hatespeech By public relations professionals. *Source*: Expert survey 2019 and 2022.

indicated that reporting to platforms under community standards had significantly increased over time ($p = 0.08$); more PR professionals relied on content moderation, such as hiding comments ($p = 0.02$) or deleting comments ($p = 0.01$). Overall, user complaint mechanisms were only used by a small minority of expert users with a noticeable but slight increase over time ($p = 0.03$). We comment on how these patterns may compare to private users in the conclusion.

Role of organizations

What is the relationship between the organization's sector and the choice of counterstrategies among PR professionals? Due to a significantly smaller number of respondents who completed the questionnaire in 2022 during the Covid pandemic, our analysis of this second research question is limited to the 2019 survey, which allows us to obtain substantively and statistically significant results due to its larger sample size.

Reporting under NetzDG was coded as a dummy variable, so we use multivariate probit regression. Since we asked respondents to report counterstrategies of the last incident of online hatespeech, the analysis is conducted only among those who had experienced online hatespeech in the past (i.e., 45% of respondents). To investigate our second research question, we estimate the relationship between the different sectors and counterstrategies, resulting in the following model:

$$Counterstrategy = \alpha + \beta_1\, public\ sector + \beta_2\, civil\ society + \beta_3\, controls + u.$$

Whether differences across sectors play a role in explaining the choice of counterstrategies hinges on the magnitude of coefficients for β_1 and β_2. Business serves as the baseline for comparison.

We first explore the role of sectors in explaining the choice of reporting. Table 1 and Figure 3 display the result of our first model on *Prior reporting mechanisms* and *Reporting under NetzDG*. Public sector institutions were more likely to report users to the platform or to file a criminal complaint for incitement of hatespeech compared to the two other sectors. This result is significant at the 95% level and positively correlated with *Prior reporting mechanisms*. However, for *Reporting under NetzDG* public sector institutions were similar to other types of organizations and equally unlikely to choose NetzDG's reporting mechanisms. This indicates that PR professionals—no matter which sector—were unlikely to proactively resort to new regulatory options. Next, we turn to content moderation and intervention. We only report dependent variables with substantively and statistically significant findings. As displayed in Table 2 and Figure 4, civil society organizations were particularly likely to hide comments ($\beta_2 = 0.\ 99$) in comparison to the other two sectors, although public sector institutions also relied on this content moderation strategy ($\beta_1 = 0.72$). Intervention and counterspeech were particularly likely to be practiced among civil society organizations and political parties. Significant at the 99% level, they were highly likely to engage in counterspeech and mobilize supporters ($\beta_2 = 0.86$ and 1.22, respectively).

In comparison, the nonsignificant estimate for businesses indicates they vary in terms of their strategies in countering online hatespeech without leaning toward any particular one. However, when making slight adjustments to the model we found that the business sector had a positive, strong, and significant predicted probability to deactivate comments as a strategy for moderating content, as illustrated in Table 3. Instead of choosing to hide comments as in other types of organizations, PR professionals in businesses seem to opt for deletion.

TABLE 1 Multivariate probit regression results for reporting as counterstrategy.

	Prior reporting mechanisms	Reporting under NetzDG
Independent variables		
Public sector institution	0.48** (0.24)	0.04 (0.38)
Civil society organization and political parties	0.23 (0.28)	−0.14 (0.48)
Control variables		
Engagement with audience	0.13 (0.08)	0.01 (0.13)
Work related to political and societal debate	−0.03 (0.37)	0.34 (0.64)
Size of organization	0.06 (0.34)	0.65 (0.53)
Female	0.08 (0.19)	−0.42 (0.32)
Age	14.97** (7.17)	−5.05 (10.23)
Age-squared	−12.23** (5.58)	2.76 (7.99)
Organization in East Germany	0.33 (0.36)	0.57 (0.51)
Regional trend in hatespeech	0.67 (0.97)	−2.13 (1.46)
Consequences on mental health	0.11 (0.37)	1.18** (0.58)
Constant	−6.18*** (2.38)	1.15 (3.34)
Observations	241	241
Pseudo-R^2	0.06	0.11

Standard errors are given within parentheses.

Abbreviation: NetzDG, German Network Enforcement Act.

*$z < 0.1$

**$z < 0.05$

***$z < 0.01$.

Source: 2019 expert survey.

SUMMARY

Our findings reveal that reporting under NetzDG is only used by a small minority of PR professionals. Despite slight increases in usage over time—possibly in response to improvements in user-friendliness after amendments to NetzDG in 2021—only roughly 10% of experts chose this counterstrategy when encountering hatespeech on social media. Our second goal was to explore the role of the type of organization in choosing to report under NetzDG as opposed to alternative strategies. We found that PR experts working at public sector institutions were particularly likely to report online hatespeech under pre-existing reporting mechanisms. This is consistent with research on institutions, which has detected strong evidence that organizations tend to not necessarily go beyond the established tools and tend to adapt existing strategies when facing new issues (Sydow et al., 2009). In addition to pre-existing reporting mechanisms, PR professionals relied on content moderation or forms of intervention. Civil society organizations and political parties opted for counterspeech and mobilizing supporters. Businesses deploy a wide range of strategies. These findings are significant because organizations play an important role in policy implementation. Complaint mechanisms are most compatible with pre-existing strategies among public sector institutions, but unlikely to be used by organizations that are motivated to engage and build community. As a result, civil society organizations, political parties, or businesses are

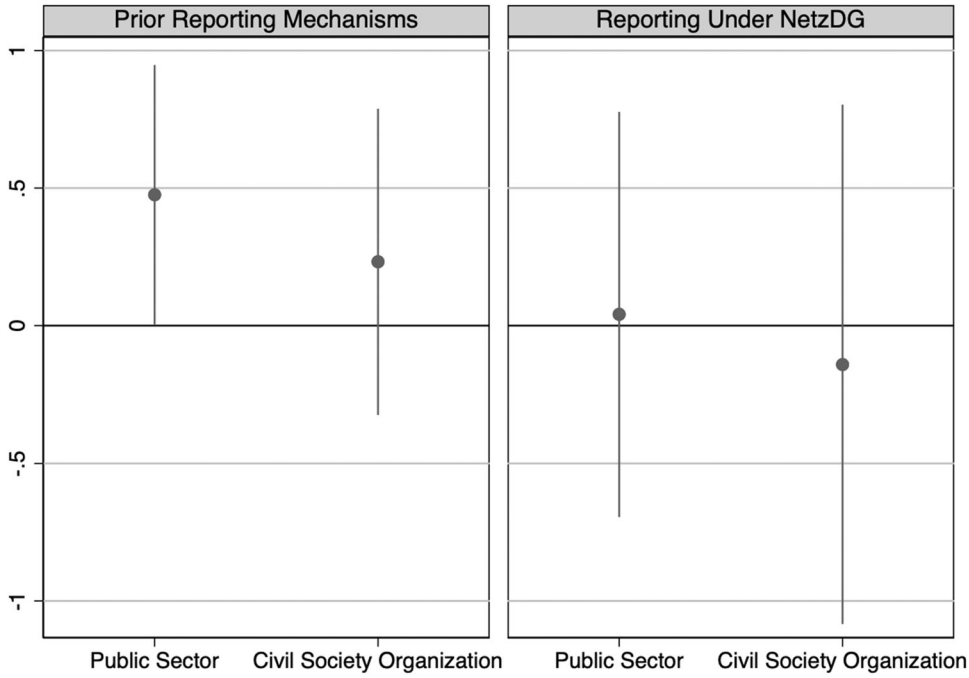

FIGURE 3 Predicted probability of public sector and civil service organizations/political parties to report online hate speech. *Source*: Expert survey 2019.

unlikely the organizations to place significant pressure on platforms to change internal processes to improve the civil discussion.

LIMITATIONS

Obviously, these findings are limited in terms of generalizability. The German Federation of All PR professionals (*Bundesverband Deutscher Pressesprecher*) has the largest and nearly complete database, but the low response rate poses limitations on our ability to generalize. However, representativeness is not the main aim of expert surveys, which attempt to uncover relationships between factors that may explain systematic variation in judgment among experts. As explained earlier, expert surveys contained enough variation to examine the relationship between organization sector and strategies to combat online hatespeech.

A second limitation relates to the lack of comparable data *before* the implementation of NetzDG, which would allow us to draw more formative conclusions about causality. However, before NetzDG came into effect in January 2018 PR professionals had no reporting options under NetzDG. It is thus reasonable to compare survey results against no use of NetzDG reporting options.

A final limitation relates to understanding how the technological design of platforms may lead users to use some platforms more compared to others. Our dependent variable was designed and pretested based on qualitative interviews with the specific goal to develop a comprehensive list of options that is available to all organizations independent of sector or usage of specific platforms. We also tested whether specific platforms facilitate the usage of

TABLE 2 Multivariate probit regression results for hiding comments, counterspeech, and mobilizing supporters as counterstrategies.

Independent variables	Hide comments	Counterspeech	Mobilizing supporters
Public sector institution	0.72*** (0.26)	0.22 (0.25)	0.62* (0.32)
Civil society organization and political parties	0.99*** (0.32)	0.86*** (0.29)	1.22***(0.34)
Engagemenet with audience	0.21** (0.08)	0.19** (0.08)	0.18* (0.10)
Work related to political and societal debate	−0.43 (0.39)	−0.06 (0.38)	0.23 (0.44)
Size of organization	1.03*** (0.37)	0.80** (0.35)	0.06 (0.40)
Female	−0.33* (0.20)	−0.16 (0.20)	0.07 (0.23)
Age	−11.59* (6.01)	−13.29** (6.26)	−3.15 (7.51)
Age-squared	8.18* (4.62)	9.02* (4.85)	3.05 (5.70)
Organization in East Germany	0.54 (0.36)	−0.28 (0.43)	0.18 (0.43)
Regional trend in hatespeech	0.76 (1.00)	−2.15** (0.99)	−1.31 (1.21)
Consequences on mental health	0.55 (0.37)	0.60 (0.38)	1.14***(0.42)
Constant	1.38 (2.03)	4.44** (2.15)	−1.07 (2.63)
Observations	241	241	241
Pseudo-R^2	0.11	0.11	0.17

Standard errors are given within parentheses.

*$z < 0.1$

**$z < 0.05$

***$z < 0.01$.

Source: 2019 expert survey.

certain strategies. When adding dummy variables for using Instagram, Facebook, Twitter, YouTube, and LinkedIn, results remain substantively stable.

CONCLUSION

Our findings have implications for policy and scholarly debates surrounding social media governance. First, process-based approaches to social media governance aim to change platforms' internal procedures in an attempt to introduce a greater emphasis on public interest, stressing a need for greater transparency, public responsibility, and accountability. While it may be desirable to ensure fairness and citizens' rights for normative reasons (Bayer, 2022), user-complaint mechanisms are only effective if citizens make use of new reporting options. This study demonstrates that organizations are unlikely to increase pressure on technology companies via user-complaint mechanisms. More research is needed to compare organizational communication strategies to strategies by private users. Transparency reports by platforms indicate that less content is removed after complaints are made under NetzDG compared to complaints about community guidelines (see, e.g., Google, 2022). To allow for better comparison transparency reports for NetzDG and the DSA should require platforms to not only report numbers of complaints by individual users but also numbers of complaints by users under community guidelines among users within the jurisdiction of the respective regulation. This would help to provide needed

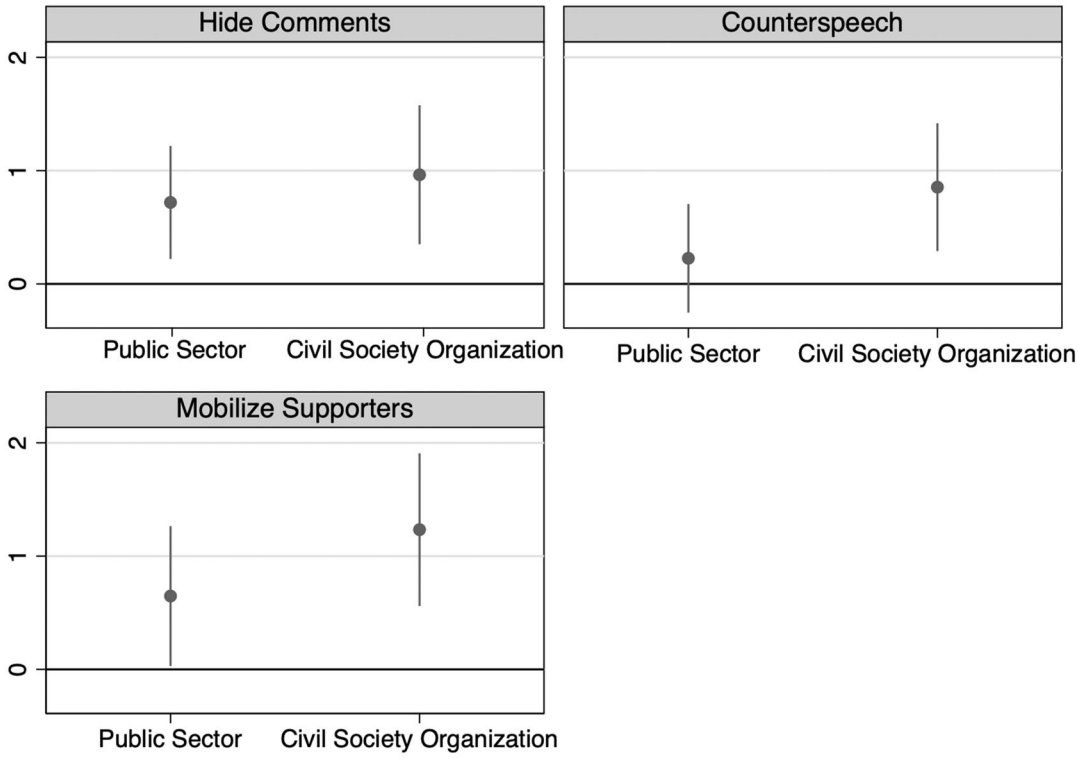

FIGURE 4 Predicted probability of public sector and civil service organizations/political parties to intervene. *Source*: Expert survey 2019.

information on the usage of user-complaint mechanisms in comparison to other reporting options.

Second, our findings expand current discussions on content moderation to include the professional world. While existing research on content moderation concentrates on the state, social media companies or users as a general category (Einwiller & Kim, 2020; He, 2020; Konikoff, 2021; Riedl et al., 2021), further research is necessary regarding organizational communication, building on more general insights on sectoral differences in communication practices (see, e.g., Einwiller & Kim, 2020; Liu et al., 2010). Future research may further investigate the reasons for choosing specific counterstrategies regarding online hatespeech. Some interviewees explained that decisions to opt for reporting mechanisms constitute a change in strategy among the organization, which required resources. Other reasons for not choosing to report may relate to concerns about the organization being seen as complicit in policing online activity. More research is needed to bring to light the rationale for the specific strategic choices made.

Finally, more research is needed beyond procedural-based and content moderation approaches toward combating online hatespeech. A number of studies point toward innovative policy instruments that have so far received less attention. Shklovski and Némethy (2023) outline a number of strategies to encourage engineers at tech companies to integrate ethical considerations in their everyday work practices, such as changes in documentation, testing, certification programs, and forms of oversight. Others argue in favor of a more radical change of the business model, which is seen as a root cause for amplifying fake news, extremism, polarization, and online hatespeech (Bennett, 2023; Landwehr

TABLE 3 Multivariate probit regression results for deactivating comments as counterstrategy.

	Deactivating comments
Independent variables	
Company	1.13** (0.57)
Control variables	
Engagement with audience	0.08 (0.21)
Work related to political and societal debate	0.39 (0.85)
Size of organization	−1.80** (0.85)
Female	0.28 (0.45)
Age	5.87 (13.89)
Age-squared	−4.06 (10.82)
Regional trend in hatespeech	2.37 (2.64)
Consequences on mental health	−0.43 (0.97)
Constant	−6.21 (5.10)
Observations	241
Pseudo-R^2	0.19

Standard errors are given within parentheses.

*$z < 0.1$

*** $z < 0.01$

**$z < 0.05$.

Source: 2019 Expert survey.

et al., 2023). Our findings indicate that such alternative options are worth closer investigation when attempting to reduce hatespeech on social media.

ACKNOWLEDGMENTS

This research was conducted in collaboration with the Amadeu-Antonio Stiftung and Bundesverband deutscher Pressesprecher e.V. with funding from the German Ministry for Family, Senior Citizens, Women, and Youth (Bundesministerium für Familie, Senioren, Frauen und Jugend, BMFSFJ). For continuous support even beyond the funding period, the authors are grateful to Regine Kreitz, Ronny Fechner, and Jürgen Kornmann. We are also grateful for comments on earlier drafts by the participants of the workshop on hatespeech on social media at the Hertie School, co-organized with Simon Munzert. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Daniela Stockmann http://orcid.org/0000-0003-0998-2752

ENDNOTES

[1] Since representativeness is not the main aim of expert surveys, they tend to have low response rates (Schwadel & Dougherty, 2010). The 2019 expert survey was conducted between April 26, 2019 and May 24,

2019, resulting in a total of 711 respondents, with 566 respondents completing the questionnaire, leading to a response rate of 2.84%. The 2022 expert survey was conducted between February 28, 2022 and March 31, 2022, resulting in a total of 544 respondents and 363 completed questionnaires with a response rate of 3.03%.

[2] Nonresponses to the dependent variable are low in both years (2019: $n = 7$; 2022: $n = 3$).

REFERENCES

Alford, J. (1998). A public management road less travelled: Clients as co-producers of public services. *Australian Journal of Public Administration*, *57*(4), 128–137. https://doi.org/10.1111/j.1467-8500.1998.tb01568.x

Alford, J. (2002). Why do public-sector clients coproduce?: Toward a contingency theory. *Administration & Society*, *34*(1), 32–56. https://doi.org/10.1177/0095399702034001004

Balkin, J. M. (2018). Free speech is a triangle. *Columbia Law Review*, *118*(7), 2011–2056.

Bayer, J. (2022). Procedural rights as safeguard for human rights in platform regulation. *Policy & Internet*, *14*(4), 755–771. https://doi.org/10.1002/poi3.298

Bayer, J., Holznagel, B., Korpisaari, P., & Woods, L. (2021). *Perspectives on platform regulation: Concepts and models of social media governance across the globe* (Vol. 1, 1st ed.). Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748929789

Bennett, W. L. (2023). Killing the golden goose? A framework for regulating disruptive technologies. *Information, Communication & Society*, *26*(1), 16–36. https://doi.org/10.1080/1369118X.2021.1994625

Beutel, A., Weine, S. M., Saeed, A., Mihajlovic, A. S., Stone, A., Beahrs, J. O., & Shanfield, S. B. (2016). Guiding principles for countering and displacing extremist narratives. *Journal of Terrorism Research*, *7*(3), 35–49. https://doi.org/10.15664/jtr.1220

Bloch-Wehba, H. (2020). Automation in moderation. *Cornell International Law Review*, *53*, 41–96.

Brudney, J. L. (1987). Coproduction and privatization: Exploring the relationship and its implications. *Journal of Voluntary Action Research*, *16*(3), 11–21. https://doi.org/10.1177/089976408701600303

Brudney, J. L., & England, R. E. (1983). Toward a definition of the coproduction concept. *Public Administration Review*, *43*(1), 59–65. https://doi.org/10.2307/975300

Bundesministerium der Justiz. (2017). Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz—NetzDG) [Act to improve law enforcement in social networks (Network Enforcement Act—NetzDG)]. *Bundesgesetzblatt Jahrgang 2017*, *1*(61), 3352–3355.

Bundesministerium der Justiz. (2021). Gesetz zur Änderung des Netzwerkdurchsetzungsgesetzes [Law amending the Network Enforcement Act]. *Bundesgesetzblatt Jahrgang 2021*, *1*(29), 1436–1443.

Davies, G., Neudecker, C., Ouellet, M., Bouchard, M., & Ducol, B. (2016). Toward a framework understanding of online programs for countering violent extremism. *Journal for Deradicalization*, *6*, 51–86.

Díaz, Á., & Hecht-Felella, L. (2021). *Double standards in social media content moderation* (pp. 1–40). Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation.

DiFranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018). Upstanding by design: Bystander intervention in cyberbullying. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (1–12). https://doi.org/10.1145/3173574.3173785

Douek, E. (2021). Governing online speech: From "Posts-As-Trumps" to proportionality and probability. *Columbia Law Review*, *121*(3), 759–834.

Eifert, M., von Landenberg-Roberg, M., Theß, S., & Nora, W. (2020). *Evaluation des NetzDG Im Auftrag des BMJV [Evaluation of the NetzDG on behalf of the BMJV]*. Bundesministerium der Justiz. https://www.bmj.de/SharedDocs/Downloads/DE/News/PM/090920_Juristisches_Gutachten_Netz.pdf?__blob=publicationFile&v=1

Einwiller, S. A., & Kim, S. (2020). How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy & Internet*, *12*(2), 184–206. https://doi.org/10.1002/poi3.239

Ganesh, B., & Bright, J. (2020). Countering extremists on social media: Challenges for strategic communication and content moderation. *Policy & Internet*, *12*(1), 6–19. https://doi.org/10.1002/poi3.236

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

Google. (2022). *Removals under the Network Enforcement Law*. Google Transparency Report. https://transparencyreport.google.com/netzdg/youtube

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany's NetzDG. *Telecommunications Policy*, *46*(9), 102411. https://doi.org/10.1016/j.telpol.2022.102411

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(50), e2116310118. https://doi.org/10.1073/pnas.2116310118

Hassel, A. (2010). Multi-level governance and organized interests. In H. Enderlein, S. Wälti, & M. Zürn (Eds.), *Handbook on multi-level governance* (pp. 153–167). Elgar.

He, D. (2020). Governing hate content online: How the Rechtsstaat shaped the policy discourse on the NetzDG in Germany. *International Journal of Communication*, *14*, 3746–3768.

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2), 1–18. https://doi.org/10.14763/2019.2.1398

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science*, *22*(1), 93–109. https://doi.org/10.1146/annurev-polisci-051517-012343

Howlett, M. (2019). Organizational implementation tools. In M. Howlett & M. Howlett (Eds.), *Designing public policies: Principles and instruments* (2nd ed., pp. 165–210). Routledge.

Huang, B., & Yu, J. (2019). Leading digital technologies for coproduction: The case of "visit once" administrative service reform in Zhejiang Province, China. *Journal of Chinese Political Science*, *24*(3), 513–532. https://doi.org/10.1007/s11366-019-09627-w

Jones, P., Speed, C., & Lorne, C. (2016). Using technology to help communities shout louder. In L. Richardson & C. Durose (Eds.), *Designing public policy for co-production: Theory, practice, and change*. Policy Press.

Klausa, T. (2023). Graduating from 'new-school'—Germany's procedural approach to regulating online discourse. *Information, Communication & Society*, *26*(1), 54–69. https://doi.org/10.1080/1369118X.2021.2020321

Kleinsteuber, H. J. (2004). The Internet between regulation and governance. In C. Moeller & A. Amouroux (Eds.), *The media freedom internet cookbook* (pp. 61–75). Organization for Security and Cooperation in Europe (OSCE).

Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, *13*(4), 502–521. https://doi.org/10.1002/poi3.265

Kraft, M. E., & Furlong, S. R. (2010). *Public policy: Politics, analysis, and alternatives* (3rd ed.). CQ Press.

Kraft, M. E., & Furlong, S. R. (2013). Chapter 2: Government institutions and policy actors. In M. E. Kraft & S. R. Furlong (Eds.), *Public policy: Politics, analysis, and alternatives* (6th ed., pp. 82–138). CQ Press.

Kuczynski, P.-P. (1999). Privatization and the private sector. *World Development*, *27*(1), 215–224. https://doi.org/10.1016/S0305-750X(98)00123-5

Landwehr, M., Borning, A., & Wulf, V. (2023). Problems with surveillance capitalism and possible alternatives for IT infrastructure. *Information, Communication & Society*, *26*(1), 70–85. https://doi.org/10.1080/1369118X.2021.2014548

Liesching, M. (2020, June15). *Stellungnahme zum Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes [Statement on the draft law amending the Network Enforcement Act]* [Letter to Deutscher Bundestag]. https://www.bundestag.de/resource/blob/700788/83b06f596a5e729ef69348849777b045/liesching-data.pdf

Liu, B. F., Horsley, J. S., & Levenshus, A. B. (2010). Government and corporate communication practices: Do the differences matter? *Journal of Applied Communication Research*, *38*(2), 189–213. https://doi.org/10.1080/00909881003639528

Lodge, M., & Wegrich, K. (2012). *Managing regulation regulatory analysis, politics and policy*. Palgrave Macmillan.

Maestas, C. (2018). Expert surveys as a measurement tool: Challenges and new frontiers. In L. R. Atkeson & R. M. Alvarez (Eds.), *The Oxford Handbook of polling and survey methods* (pp. 583-608). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190213299.013.13

Mair, J., & Seelos, C. (2021). Organizations, social problems, and system change: Invigorating the third mandate of organizational research. *Organization Theory*, *2*(4), 26317877211054858. https://doi.org/10.1177/26317877211054858

Marwick, A., & Miller, R. (2014). *Online harassment, defamation, and hateful speech: A primer of the legal landscape No. 2; Fordham Center on Law and Information Policy Report* (pp. 1–74). Fordham Law School. https://ir.lawnet.fordham.edu/clip/3

Massey, C. R. (1992). Hate speech, cultural diversity, and the foundational paradigms of free expression. *UCLA Law Review*, *40*(103), 103–197.

Olson, M. (1971). *The Logic of Collective Action: Public Goods and the Theory of Groups* (Vol. 124). Harvard University Press.

Ostrom, E. (1996). Crossing the great divide: Coproduction, synergy, and development. *World Development*, *24*(6), 1073–1087. https://doi.org/10.1016/0305-750X(96)00023-X

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: rethinking regulation and responses* (pp. 37–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.006

Patton, M. Q. (2008). *Utilization-Focused Evaluation: The New Century Text* (4th ed.). SAGE Publications.

Patton, M. Q. (2012). *Essentialsof Utilization-Focused Evaluation*. SAGE Publications, Inc.

Perry, J. L., & Rainey, H. G. (1988). The public–private distinction in organization theory: A critique and research strategy. *The Academy of Management Review*, *13*(2), 182–201. https://doi.org/10.2307/258571

Persily, N. & Tucker, J. A. (Eds.). (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press. https://doi.org/10.1017/9781108890960

Poocharoen, O., & Ting, B. (2015). Collaboration, co-production, networks: Convergence of theories. *Public Management Review*, *17*(4), 587–614. https://doi.org/10.1080/14719037.2013.866479

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*, *13*(3), 433–451. https://doi.org/10.1002/poi3.257

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, *22*(1), 1–15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T

Schwadel, P., & Dougherty, K. D. (2010). Assessing key informant methodology in congregational research. *Review of Religious Research*, *51*(4), 366–379.

Sellars, A. (2016). *Defining hate speech* (no. 2016–20; Berkman Klein Center Research Publication). Berkman Klein Center. https://doi.org/10.2139/ssrn.2882244

Shklovski, I., & Némethy, C. (2023). Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society*, *26*(1), 37–53. https://doi.org/10.1080/1369118X.2021.2014547

Steen, T., & Tuurnas, S. (2018). The Roles of the Professional in Co-Production and Co-Creation Processes. In T. Brandsen, B. Verschuere, & T. Steen (Eds.), *Co-Production and Co-Creation: Engaging Citizens in Public Services* (1st ed., pp. 80–92). Routledge.

Sutton, R. (1999). *The policy process: An overview* [ODI Working Paper No. 118]. Overseas Development Institute. https://cdn.odi.org/media/documents/2535.pdf

Sydow, J., Schreyögg, G., & Koch, J. (2009). Organizational path dependence: Opening the black box. *Academy of Management Review*, *34*(4), 689–709. https://doi.org/10.5465/amr.34.4.zok689

Szmania, S., & Fincher, P. (2017). Countering violent extremism online and offline. *Criminology & Public Policy*, *16*(1), 119–125. https://doi.org/10.1111/1745-9133.12267

Taylor, A., Lee, H. R., Kubota, A., & Riek, L. D. (2019). Coordinating clinical teams: Using robots to empower nurses to stop the line. *Proceedings of the ACM on Human–Computer Interaction*, *3*(CSCW), 1–30. https://doi.org/10.1145/3359220

Thomas, J. C., Box, R. C., & Rimmerman, C. A. (1999). Bringing the public into public administration: The struggle continues. *Public Administration Review*, *59*(1), 83–88. https://doi.org/10.2307/977482

Tworek, H., & Leerssen, P. (2019). *An analysis of Germany's NetzDG Law* (Transatlantic High Level Working Group on content moderation online and freedom of expression). Institute for Information Law (IViR). https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

Van Eerten, J.-J., & Doosje, B. (2019). *Challenging extremist views on social media: Developing a counter-messaging response* (1st ed.). Routledge.

van Eijk, C., & Gascó, M. (2018). Unravelling the co-producers: Who are they and what motivations do they have? In T. Brandsen, B. Verschuere, & T. Steen (Eds.), *Co-production and co-creation* (1st ed., pp. 63–76). Routledge.

Vaughan, S. K., & Arsneault, S. (2013). *Managing Nonprofit Organizations in a Policy World* (1st ed.). CQ Press.

Vermeulen, M. (2019). *Online content: To regulate or not to regulate—Is that the question?* [APC Issue Paper]. Association for Progressive Communications. https://www.apc.org/sites/default/files/OnlineContentToRegulateOrNotToRegulate.pdf

Warren, R., Rosentraub, M. S., & Harlow, K. S. (1984). Coproduction, equity, and the distribution of safety. *Urban Affairs Quarterly*, *19*(4), 447–464. https://doi.org/10.1177/004208168401900404

How to cite this article: Stockmann, D., Schlosser, S., & Ksatryo, P. (2023). Social media governance and strategies to combat online hatespeech in Germany. *Policy & Internet*, *15*, 627–645. https://doi.org/10.1002/poi3.348