The 2021 German Federal Election on Social Media: An Analysis of Systemic Electoral Risks Created by Twitter and Facebook Based on the Proposed EU Digital Services Act

Report

Authors

Johanne Kübler, Marie-Therese Sekwenz, Felicitas Rachinger, Anna König, Rita Gsenger, Eliška Pírková, Ben Wagner, Matthias C. Kettemann, Michael Krennerich, Carolina Ferro

August 2021

`

`

Contents

`

Executive summary

Safeguarding democratic elections is hard. Social media plays a vital role in the discourse around elections and during electoral campaigns. Social media platforms have become central spaces for electoral campaigns, often substituting traditional media outlets. Many politicians and parties communicate their messages primarily on their Twitter and Facebook profiles. In that regard, these platforms can be a valuable tool, but they also contribute to risks such as the dissemination of disinformation or other content that can infringe the right to free and fair elections.

This study provides a risk assessment of the 'systemic electoral risks' created by Twitter and Facebook and the mitigation strategies employed by the platforms. It is based on the 2020 proposal by the European Commission for the new Digital Services Act (DSA) in the context of the 2021 German federal elections. Therefore, **this study provides an external risk assessment** regarding the right to free and fair elections on very large online platforms (VLOPs), focusing on Twitter and Facebook and their roles during the German federal elections that will take place on 26 September 2021. The data collection period covered the second half of May 2021.

We analysed three systemic electoral risk categories: 1) the dissemination of illegal content, 2) negative effects on electoral rights, and 3) the influence of disinformation. In this context, the present study found a significant number of problematic posts and tweets during the analysis, with **6.72% of all election-related Facebook posts** and **5.63% of election-related tweets** falling into at least one of the risk categories in our codebook, meaning they were potentially illegal, disinformation, or infringements of electoral rights.

Of the problematic posts on Facebook, 4.05% were likely illegal under German law, 35.14% violated the platform's community standards or Terms of Service, 46,65% were violations of electoral rights, and 93.24%

`

could be considered disinformation. Similarly, for the Twitter sample, of the problematic tweets, 14.52% broke platform rules, 51.61% infringed on electoral rights, and 100% were considered disinformation.

The key policy recommendations we developed as a result are as follows:

- Platforms need to create more effective and sustainable response mechanisms to do more to safeguard elections.
- Platforms' responses to problematic content should be based on rigorous scholarly research.
- Platforms have to become more transparent about content moderation tools they deploy, including algorithmic transparency.
- Platforms' Terms of Service need to be expanded to cover all forms of disinformation and electoral rights more effectively, especially in times of elections.
- Responses to problematic content should be harmonised across platforms.
- Limiting mitigation measures on illegal content is ineffective to safeguard elections, as almost all the observed problematic content is legal.
- Platforms should focus on content curation and moderation and design measures that promote free expression and user agency.
- There is a need for more clearly delineated and easier to reproduce categories of analysis; thus, additional research and policy development are needed to operationalise and clearly delineate what constitutes disinformation and electoral rights violations.
- DSA risk assessment methodology should be expanded beyond very large online platforms.
- Access to platform data needs to be easier for researchers for them to hold platforms accountable.
- Public auditing intermediaries should be introduced to further secure and strengthen the independence of auditors and the auditing regime.

`

Acknowledgements

The authors wish to thank Felix Kartte and the Luminate Group’s team for their great partnership and for providing the funding for this project. Also, we would like to acknowledge the valuable contributions of all our interview participants.

`

Introduction

Safeguarding democratic elections is hard. Although frequently taken for granted, it is so central to democratic governance, and yet so difficult to ensure. Safeguarding democratic elections is not simply about ensuring that votes are counted correctly. The media environment around elections also plays a critical role. As the media environment has been changing rapidly in the past decades, the risks of free and fair elections are also evolving rapidly.

Social media plays a central role in these media environments in many parts of the world. This study investigates in detail the systemic risks posed by social media in the context of elections, and the ways in which these risks can be mitigated. It also assesses the extent to which social media platforms, such as Facebook and Twitter, sufficiently reduce these risks or whether they could be doing more.

In this context, the 2020 proposal of the European Commission for the new Digital Services Act (DSA) is an important piece of legislation that promises to significantly strengthen the European accountability regime for online platforms. Article 26 of the DSA forces very large online platforms (VLOPs)[1] to identify significant systemic risks stemming from the operation of their platforms, and Article 27 proposes mitigation measures that these platforms should implement. However, what would a concrete risk assessment of an online platform based on the proposed DSA look like in practice?

To address this question, this study proposed to conduct an **external risk assessment** without access to internal platform data. Of course, such risk assessments are very difficult. Consequently, the study was only able to conduct a much smaller version of a risk assessment than would actually be legally necessary under the proposed DSA. However, we believe our

[1] 'Very large online platforms' are defined by the DSA as those having more than 45 million recipients of the service, which is the equivalent of 10% of the European Union's population (European Commission, 2020, Art. 25).

`

work can serve as an initial demonstration of what such risk assessments for electoral processes could look like, and contribute to the debate on how to implement them in practice.

The risk assessment was carried out in the context of the German federal elections that will take place on 26 September 2021, taking into consideration the two large online platforms mentioned previously: Twitter and Facebook. The study focused exclusively on '**systemic electoral risks'** rather than examining other areas of systemic risks also raised by large online platforms.

The risk assessment and the analysis provided by this study include:

- A study of systemic electoral risks on two very large online platforms (VLOPs)—Facebook and Twitter—which focused on three categories: the dissemination of illegal content, negative effects on electoral rights, and the influence of disinformation;
- A theoretical framework and innovative methodological approach that enables an external assessment of systemic electoral risks in online platforms;
- A codebook developed for analysis according to these three systemic risk categories;
- An analysis of 2202 Facebook and Twitter posts;
- A report of Twitter and Facebook's contributions to systemic risks concerning the right to free and fair elections;
- A presentation of strategies to mitigate risks concerning the right to free and fair elections by Twitter and Facebook;
- The inclusion of assessments by experts from various fields;
- A list of policy recommendations in light of the DSA.

Lastly, it should be noted that the whole of the EU DSA remains a legislative proposal, including Article 26 on 'risk assessment' and Article 27 on 'mitigation of risks'. As such, we hope that our experience in conducting a concrete risk assessment in practice can contribute to further development of the DSA. Thus, we trust that this report can

contribute to a better understanding of the degree to which the DSA is effective in safeguarding European elections and where more still needs to be done.

`

1. Systemic electoral risks: Theoretical and methodological approach

Online platforms, such as social media sites, online marketplaces, communities, and forums, have become major avenues for the distribution of information and debates on politics, especially in the context of elections. These online platforms abound in text, images, videos, posts, tweets, and reviews created and shared organically by users. Furthermore, political actors are increasingly using online platforms to engage with the electorate, as well as investing in political advertising on these platforms, using the tools provided to advertisers to micro-target the electorate.

Misinformation, disinformation, and propaganda tactics are not unique to our era (Ireton, Posetti, and UNESCO 2018, 15). However, the near universal adoption of internet-based technologies has amplified the impact of these tactics on political events. Politicians and other political actors increasingly use internet-based communication, especially online platforms, to engage with voters directly. This move towards internet-based communication is reflected in the changing media consumption habits of the population in eight surveyed Western European countries (Matsa 2018). According to this survey, although television remains the most important news source, online news consumption comes second, and in two countries, Sweden and Denmark, it surpasses television as the primary news source. The result of these developments is the demise of traditional quality media, the gradual erosion of editorial standards, and increasingly sensationalised news coverage (Anand 2017).

The 2016 US presidential election, which installed Donald J. Trump as President, and the UK's referendum decision to leave the European Union ('Brexit') the same year were watershed moments for the public perception of the online platforms' role in elections. Given that pollsters and traditional media predicted a win for Hillary Clinton and a victory for the 'Remain campaign' in the UK, the public questioned the influence of

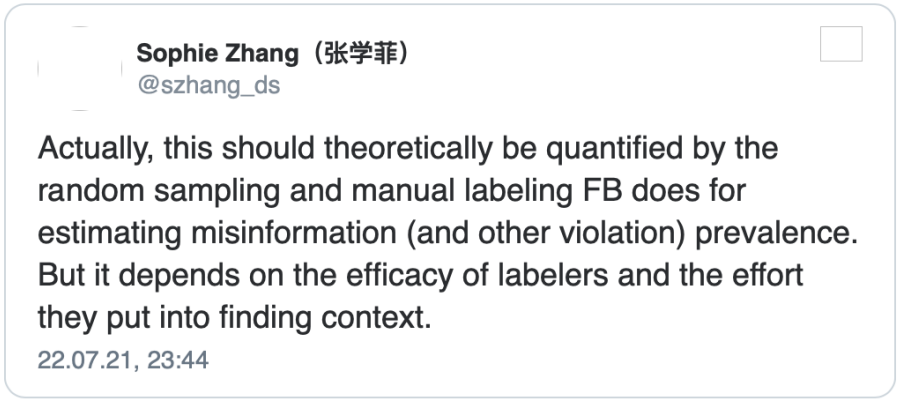
`

online platforms on the campaigns (Isaac 2016). In fact, researchers subsequently found that during the US presidential election, 25% of tweets containing a link to news outlets spread either fake or extremely biased news (Bovet and Makse 2019). Another study on the use of political bots during the UK referendum found, based on a sample of more than 1.5 million tweets, that less than 1% of sampled accounts generated almost a third of all messages, making the role of bots during Brexit small but strategic (Howard and Kollanyi 2016).

Whereas online platforms initially rejected holding any responsibility for the content published on their sites, they subsequently established a combination of human-driven and automated editorial processes to promote or remove certain content types. The so-called ‘content moderation’ is the systematic practice of a social media platform of screening content to ensure compliance with community guidelines, user agreements, laws, and regulations, and norms of appropriateness for a certain locality and its cultural context (Roberts 2017).

As a result, online platforms have sought to curb the spread of harmful content beyond the manifestly illegal—such as content in violation of the Terms of Service (ToS) or community guidelines—in numerous ways. Strategies to combat the spread of harmful content range from fact-checking suspicious information, the deletion or reduction of the reach of suspicious profiles, modifying the rules or prohibiting political ads, and design choices, such as alerting users that they might unwittingly be sharing disinformation.

As noted by former Facebook staff members and whistleblower Sophie Zhang, using a solid methodological framework and sampling strategy is crucial in this context to ensure that the resulting data can be meaningfully interpreted:



In her tweet, Zhang confirms that “one of the greatest complexities of modern social media content moderation that stymies attempts to automate it is that context and intent matters.” (Leetaru 2019). Content moderation demands discerning the complex nuance of context and intent behind a given post/tweet, which is a difficult task, with a lot of it depending on human interpretation. Therefore, the work of labellers and moderator teams is crucial, and companies need to invest in them and properly support their efforts to efficiently identify and mitigate systemic risks to their services.

To respond to the need to examine and curate a large amount of data, online platforms have developed **content moderation** systems. For instance, the main strategies regarding content moderation that Facebook and Twitter have employed over the years are:

- Fact-checking: This has become more prevalent in modern journalism, shown by the increasing numbers of fact-checking organisations being established around the world, and the creation of dedicated sections in many established media outlets. The goal is to verify the facts presented in a social media post and to produce an accurate analysis of public statements to ‘correct’ public misperceptions and increase knowledge of important issues. It is the most widespread strategy for countering disinformation on social media platforms.

`

- Deletion of content: Online platforms rely heavily on the removal of harmful and otherwise undesirable content, increasing concerns regarding its impact on the freedom of expression and information and digital rights of individuals.
- Ban/suspend user accounts after repeated offence: The so-called 'deplatforming' means the removal of a user account from a platform due to infringement of the platform rules (Rogers 2020).

These content moderation systems rely heavily on the removal of harmful and, otherwise, undesirable content. However, there are growing concerns regarding the impact of these platforms' decisions on human rights and individuals' freedom of expression and information. Many community-led platforms[2] offer alternatives to these challenges, as a previous study highlighted (Wagner et al. 2021). For instance, as an alternative to deleting undesirable content, some community-led platforms use systems that enable users to downvote/upvote content and/or other users. "While each site uses a slightly different reputation system, they generally track the behaviour of members by giving users "karma" points for their posts and other activities, as well as the ability to upvote (and, usually, also downvote) other's contributions. When a post is upvoted or downvoted by fellow members of a community, the poster receives or loses points." (Wagner et al. 2021, 27). This method of reducing the visibility of certain content is used by platforms such as slashdot.

Apart from content moderation strategies, the **design choices** that online platforms make affect which information is available, how it is displayed, and how people communicate. A recent study showed that implementing changes in platform design to promote different forms of appropriate behaviour within specific communities may be particularly effective in getting users to change their behaviour (Wagner and Kubina 2021). More recently, some important design strategies promoted by Twitter and Facebook to curb harmful content include the following:

[2] Community-led platforms are platforms partially or entirely governed by its community of users (Wagner et al. 2021).

`

- Content warning labels
- Attaching links to trusted sources
- Hiding content behind a screen
- Asking if a user really wants to share a given post/tweet
- Providing automated feedback to users who are likely to break rules

Lastly, the **systems for selecting and displaying advertisements** on online platforms, such as social media, notably through micro-targeting audiences, can be problematic for electoral rights and safeguarding free and fair elections. Political advertisements could promote positive democratic outcomes, such as facilitating increased engagement with elections or giving people information that helps them make more informed political choices. However, as currently enacted, there is a glaring lack of legal frameworks regulating online political advertisements, which means that each platform creates its own rules.

Large social media platforms have an advertisement-driven business model. However, that kind of commercial model of targeted advertising may be problematic during election campaigns. Recognising this, Twitter decided to ban all political advertising in November 2019 before the 2020 US federal elections. “While internet advertising is incredibly powerful and very effective for commercial advertisers, that power brings significant risks to politics, where it can be used to influence votes to affect the lives of millions,” company CEO Jack Dorsey tweeted. By contrast, Facebook initially rolled out a ban on political ads but afterwards implemented certain changes, such as increasing accountability mechanisms by requiring advertisers to register and thus leave an accountability trail.

After briefly presenting the background on social media and elections, and some strategies that online platforms have been developing to curb the spread of harmful content and, therefore, protect the rights to free and fair elections, the next section aims to build a theoretical framework to

`

analyse systemic risks in the context of elections and online platforms in the European Union, based on the EU DSA proposal. More specifically, this framework aims to enable testing of the performance of VLOPs—by performing a systemic risk assessment—in the context of the German Federal elections that will take place in September 2021. To achieve this aim, an innovative methodological approach was developed, which will be presented below.

1.1. Theoretical framework

This section will discuss the theoretical framework built for this study. It will expose how the systemic risk assessment required by the DSA proposal to VLOPs was tailored to the specific risks identified in the context of elections (what we named in this study 'systemic electoral risks'), considering the negative impacts they might have on free and fair elections.

1.1.1. DSA, online platforms, and systemic risks

Social media platforms control the flow of information shared on their platforms through rules codified in their algorithms. These platforms choose to promote certain content above others to keep their websites appealing to users as part of their business model. They also screen (or moderate) content to guarantee its compliance with laws and regulations, community guidelines, and user agreements.

Within the context of the new EU DSA, a draft of which was published by the European Commission in December 2020, the platforms play an important role in safeguarding fundamental rights. The role of large platforms is particularly important in the context of Article 26 of the DSA, which argues that 'very large online platforms' must take measures to prevent creating 'systemic risks.'

`

The term ‘systemic risk’ rose to prominence in discussions related to the 2008 economic crisis, when failing large financial firms with complex businesses caused ripple effects in the larger economy. Systemic risk thus describes risks that “emerge from complex system failure, where the failure of a single component leads to systemic knock-on effects” (Manheim 2020, 2).

Similarly, in the DSA proposal, the European Commission recognises that VLOPs cause significant societal risks due to the large number of recipients of the service and their role in facilitating public debate, economic transactions, and the dissemination of information, opinions, and ideas and in influencing how recipients obtain and communicate information online (European Commission 2020, Art. 53). Indeed, as the number of users of social media platforms has soared, online activity has become increasingly central to offline cultural and political events, such as the UK’s Brexit vote and the 2016 US presidential election, as mentioned previously. In this context, the online networks’ potential to splinter the public into informational echo chambers, induce ingroup/outgroup hostilities, and make participants vulnerable to misinformation and propaganda dominated the headlines (Rhodes 2021; Spohr 2017).

This study uses this interpretation of these systemic risks as an inspiration and attempts to understand the extent to which the online platforms studied adequately address the systemic risks they create in an electoral context. Specifically, Article 26 of the DSA defines **three dimensions or categories of content** that could potentially be considered relevant for platforms when conducting systemic risk assessments:

A. “the dissemination of illegal content through their services;
B. any negative effects for the exercise of the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination; […]
C. intentional manipulation of their service, including by means of inauthentic use or automated exploitation of the service, with […]

`

> actual or foreseeable effects related to electoral processes and public security." (European Commission, 2020).

Based on the analysis of these three categories of systemic risks, this study attempts to understand the extent to which online platforms have been able to prevent these risks. Further, another relevant question in this context would be how VLOPs are expected to achieve this. In this regard, Articles 26 and 27 of the DSA also offer some suggestions.

Regarding DSA Article 26, while the text below is the original text of the DSA, the structure presented is by the authors:

"Very large online platforms shall identify, analyse and assess […] any significant systemic risks […] When conducting risk assessments, very large online platforms shall take into account […] how their:

1. content moderation systems,
2. recommender systems, and
3. systems for selecting and displaying advertisement

influence any of the systemic risks […] including the potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions" (European Commission 2020).

While the three subgroups above provide some indication of how platforms could respond to systemic risks during elections, they need to be specified a little further:

1. Content moderation systems: Studying content moderation systems needs to examine both illegal content and content in violation of the ToS across all aspects of content provided by the platform. It also needs to consider both algorithmic and human content moderation. What steps do online platforms take to reduce systemic risks through content moderation?

`

2. Recommender systems/Design choices: While only recommender systems are specified by the DSA, we believe what is actually meant is the design of the platform itself. What steps do online platforms take to reduce systemic risks through the design of their platforms?

3. Systems for selecting and displaying advertisements: Electoral advertising can be extremely powerful; thus, this section is particularly important. Whether electoral ads are correctly categorised is only as important as the extent to which electoral ads make those who fund them transparent and how they are targeted at users. What steps do online platforms take to reduce systemic risks through online advertising?

Lastly, DSA Article 27 suggests, more precisely, how VLOPs should conduct risk mitigation. Paragraph 1 states:

"1. Very large online platforms shall put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 26. Such measures may include, where applicable:

(a) adapting content moderation or recommender systems, their decision-making processes, the features or functioning of their services, or their terms and conditions;

(b) targeted measures aimed at limiting the display of advertisements in association with the service they provide;

(c) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk;

(d) initiating or adjusting cooperation with trusted flaggers in accordance with Article 19;

(e) initiating or adjusting cooperation with other online platforms through the codes of conduct and the crisis protocols referred to in Article 35 and 37 respectively." (European Commission 2020).

`

Importantly, the DSA is not the first legal framework of this kind to require that online platforms ensure that they do not spread illegal content. Many countries oblige social networks to remove any content that is ‘manifestly unlawful’. EU law outlaws four types of content: (i) child sexual abuse material through the Child Sexual Abuse and Exploitation Directive (2011/93/EU); (ii) racist and xenophobic hate speech through the Counter-Racism Framework (2008/913/JHA); (iii) terrorist content through the Counter-Terrorism Directive ((EU) 2021/947); and (iv) content infringing intellectual property rights through the Copyright in Digital Single Market Directive ((EU) 2019/790). Beyond these categories, what is considered illegal content varies widely among member states.

Thus, “the same type of content may be considered illegal, legal but harmful or legal and not harmful” across EU member states (De Streel et al. 2020). Together with the Digital Markets Act, the DSA’s objective is to update the European Union’s digital regulation framework, in particular by modernising the e-Commerce Directive adopted in 2000 through a single set of new rules applicable across the entire EU, aiming to secure the protection of users’ fundamental rights online and create a stronger public oversight of online platforms.

1.1.2. Systemic electoral risks and categories for risk assessment

Article 26 of the DSA 2020 proposal considers that VLOPs “shall identify, analyse and assess, [...] any significant systemic risks stemming from the functioning and use made of their services in the Union” (European Commission 2020). However, risks in the context of elections are mentioned but not detailed in the current DSA proposal. Therefore, this study intends to tailor the risk assessment imposed by the DSA for VLOPs to the context of elections in the European Union.

`

In this study, we call this type of risk **‘systemic electoral risks,’** which refers to the impacts of systemic risks—stemming from the functioning and use of VLOPs services—on democratic elections. These systemic risks may vary from disinformation or manipulative and abusive activities, and may impact the ability to safeguard free and fair elections.

To discuss systemic electoral risks, the dimensions/categories proposed in the EU DSA were adapted for this study. Thus, in this study, systemic risks are defined as primarily falling into the following three categories:

- Dissemination of illegal content
- Negative effects on electoral rights
- Disinformation

This study developed a codebook consisting of three distinct parts corresponding to each category mentioned above (see Annex 6.1). We identified different types of systemic risks for each category—so-called subcategories—which received a specific code, covering legal clauses, classifications of infringements to electoral rights during election campaigns, and various forms of disinformation. The subcategories of illegal content are based on previous work done by the authors (Tiedeke et al. 2020).[3] The electoral rights subcategories were developed in together with Michael Krennerich for this study based on his existing work in this area. Lastly, the disinformation subcategories are based on Kapantai et al. (2021).

Below, the categories created for this study to effectuate a risk assessment regarding systemic electoral risks are explained in detail.

A. Dissemination of illegal content

Content that is shared and published on social media platforms might fall under the restrictions of speech, such as libel, incitement of hatred, or defamation. Such illegal content might also fall under the category of

[3] Part of the analysis of the content is based on categories developed in a project financed by the Leibniz Institute for Media Research | Hans-Bredow-Institut (Tiedeke et al. 2020).

`

disinformation. In this regard, its wide dissemination can influence elections and infringe on individuals' electoral rights. A prominent example is the 2016 elections in the United States, where disinformation and hateful content dominated the electoral process and reportedly influenced the election's outcome (Lapowsky 2016). Beyond manifestly illegal content, social networks remove content in contravention of their own ToS, a legal document a person must agree to abide by when registering an account.

This study designed 63 codes for the 'illegal content category' based on the comprehensive taxonomy of German national and international law developed by Tiedeke et al. (2020). Created to evaluate the quality of content governance decisions in online forums in Germany and Austria, the taxonomy also includes relevant aspects of platform ToS that can lead to the deletion of content. Given the focus of the present study, we used the categories related to German and international law, and the relevant ToS categories.

B. Negative effects on electoral rights

The Universal Declaration of Human Rights, accepted by the United Nations General Assembly in 1948, enshrines the rights and freedoms of all human beings (United Nations 1948). Recognising the important role of free and fair elections to ensure the fundamental right to a participatory government, its Article 21 states:

- Everyone has the right to take part in the government of his/her country, directly or through freely chosen representatives.
- Everyone has the right of equal access to public service in his country.
- The will of the people shall be the basis of the authority of government; this shall be expressed in periodic and genuine elections which shall be by universal and equal suffrage and shall be held by secret ballot or by equivalent free voting procedures.

`

Similarly, the right to free and fair elections is rooted in the founding values of the European Union: respect for human dignity, freedom, democracy, equality, the rule of law, and respect for human rights (European Commission 2018a). Hence, the European Commission has sought to enhance transparency, protect free and fair elections, and promote the democratic participation of all European citizens in various ways, for example, through its electoral package for the 2019 European Parliament election (Juncker 2018). Therefore, it seems sensible that category (b) proposed by Article 26 of the DSA, “any negative effects for the exercise of the fundamental rights” would include electoral rights and the right to free and fair elections.

Indeed, some content circulating on online platforms during election campaigns has been found to infringe on electoral rights. Electoral rights are primarily defined as the right to vote for and stand as a candidate; however, more broadly, the right to free and fair elections also creates obligations for the state to guarantee electoral rights and create the institutional framework for periodic and genuine, free and fair elections to take place (Inter-Parliamentary Council 1994). Whereas these responsibilities initially focused on enabling parties and candidates to freely communicate their views to the electorate, the changes to the media environment through the migration of a great deal of the political debate to online platforms has necessitated the European Union to recognise that state responsibilities go beyond the organisation of elections and monitoring the conduct of the election process to encompass the responsibility of enhancing democratic resilience to online disinformation and behavioural manipulation (European Commission 2018b).

The ‘electoral rights category’ developed by Prof. Michael Krennerich for this study captures the various dimensions by which disinformation can affect an election. This category was grouped into three overarching subcategories: procedural disinformation, disinformation on parties and candidates, and integrity of elections. The subcategories cover the entire

`

lifecycle of an election, including voter registration, voter identification, election campaign, election day, counting, and the publication of the results. In total, 20 codes were designed for the ‘electoral rights’ category.

C. Disinformation

Disinformation can be defined as the dissemination of false information with the aim of influencing public opinion, groups, or individuals serving political or economic interests. Contrary to misinformation, whose inaccuracies are unintended, disinformation is false information spread intentionally (Karlova and Fisher 2013). This information is often disseminated covertly and is intended to obscure the truth. The related term ‘fake news’, however, is a political expression used to criticise a news story or media outlet (HLG 2018). Online platforms implement different strategies to deal with disinformation.

The EU’s approach to disinformation is characterised by a primacy set on protecting freedom of expression and other rights and freedoms guaranteed under the EU Charter of Fundamental Rights (2012). The EU’s approach thus favours making the online sphere and its actors more transparent and accountable, thus making content moderation practices more transparent instead of criminalising or prohibiting disinformation as such (European Commission 2021, 1). Its main instrument has been the self-regulatory Code of Practice on Disinformation, which has been in force since October 2018. The Code was adopted by all major online platforms active in the EU and major trade associations representing the European advertising sector and is generally considered a substantial achievement. However, in 2020, the Commission’s Assessment of the Code of Practice found limitations due to its self-regulatory nature, gaps in the coverage of the Code’s commitments, and inconsistency and inadequacy in its application across platforms and Member States. Based on this assessment, the Commission published Guidelines particularly stressing the need to tackle the demonetisation of disinformation through a reform of the market for online advertisements, to commit online platforms to limit manipulative behaviour, strengthen user empowerment tools,

`

increase the transparency of political advertising, and further empower the research and fact-checking community (European Commission 2021, 2).

The ‘disinformation category’ created for this study is based on Kapantai et al.’s (2021) comprehensive literature review of disinformation taxonomies. It comprises 11 elements distinguishing the various forms disinformation takes in practice. A test run on 50 tweets and Facebook posts revealed that two subcategories were either irrelevant to our data or introduced noise, namely “biased” and “fake reviews”. Therefore, these subcategories were removed, leaving nine disinformation codes.

1.2. Research design, methodology, and case selection

Taking into consideration the three DSA categories adapted to discuss systemic electoral risks and the possibility of online platforms adopting different approaches to assess and mitigate systemic risks, this study explored the following questions:

1. In the context of elections, what would a risk assessment in VLOPs look like in practice, considering the dissemination of illegal content, negative effects on electoral rights, and disinformation?
2. What measures are VLOPs taking to reduce systemic electoral risks through content moderation, design choices, and online advertising?
3. Are these measures and the approaches taken by VLOPs to assess and mitigate systemic electoral risks effective?
4. To what extent is the current DSA proposal, especially Articles 26 and 27, effective in protecting European elections?

Answering these questions was particularly challenging as the authors did not have the same amount of data as the platforms did. We were only able to address these questions based on public information. Without inside privileged access to all relevant data, our methodology is necessarily

`

limited in scope to the data sources that we are able to access as external researchers and a full analysis by one of the VLOPs would need to be much more expansive. Notably, we were not able to access sufficient advertising data to be able to analyse the platforms systems for selecting and displaying advertisements, although we hope to be able to do so in future research projects. Nevertheless, we believe it is possible to make an initial attempt at what a credible risk analysis could look like, while acknowledging that due to our lack of all relevant data, such an attempt is necessarily incomplete.

In reports and statements, social media companies are keen to stress the effectiveness of their measures in limiting the prevalence of illegal and misleading information on their platforms. In the absence of an independent validation of these reports, however, the public and policymakers are currently unable to assess the veracity of these claims (Wagner et al. 2021). To explore the scope of illegal and misleading content, as well as content infringing on electoral rights, this study analysed empirical data from two VLOPs operating in Germany.

This study relied on a qualitative and quantitative research design. To explore the three dimensions or categories of potential threats emanating from social media platforms during the 2021 German federal elections, this study applied a multi-layered mixed-method research design combining: (1) a quantitative and qualitative analysis of organic user-generated content on selected social media platforms in the context of a major election to explore the scope of illegal content, disinformation, and content infringing on electoral rights on said platforms; (2) qualitative, semi-structured interviews with individuals familiar with the platforms, who could provide contextual information about the extent to which they were responding effectively to systemic risks, and information on design choices and advertising strategies; and (3) case studies of two VLOPs operating in Germany.

`

1.2.1. Step-by-step: How to conduct an electoral risk assessment

First, to assess the prevalence of potentially harmful content on social media platforms during election campaigns, we developed a codebook with subcategories (identified with codes) on dissemination of illegal content, infringements on electoral rights, and disinformation (Annex 6.1). Coding is a common technique for condensing data into identifiable topics. A code is a distilled topic applied to a text segment illustrating that topic. By using codes, researchers can search for topics across data and thereby identify patterns (Mihas and Odum Institute 2019, 2). Based on these coded data, the study estimated the proportion of data that matched our categories and were present across the respective platforms.

Subsequently, we collected the data samples necessary for this study on Twitter and Facebook (see section ‘Data Collection’ below). This was followed by a coding test conducted on 50 random tweets to assess the appropriateness of the subcategories, allowing for fine-tuning definitions with the coders, and assessing the intercoder reliability for each subcategory.

With subcategories and codes fine-tuned, a random representative sample of 1101 tweets and 1101 Facebook posts were coded. The coding process was done in parallel by two different groups of coders. Thus, each sample was coded twice. The data were then merged and the intercoder reliability for each subcategory discussed and, when necessary, the coding of was adjusted. In fact, Facebook uses a similar but distinct method to estimate the prevalence of misinformation and other harmful content on its platform, relying on random sampling and manual labelling, as former Facebook staff member and data scientist Sophie Zhang described in her tweet we mentioned earlier.

We also conducted nine semi-structured interviews with 10 individuals familiar with the platforms, primarily with current and former employees of the platforms and scholars who are experts on the topic, who were able to provide us with contextual information about the extent to which VLOPs

`

were responding effectively to systemic risks. Further, information on design choices and advertising strategies was collected during the interviews, with questions designed specifically for this purpose (see Annex 6.3). Lastly, we conducted case studies on Twitter and Facebook, based on extensive desk research and the information extracted from the interviews (see the 'Case Selection' section below).

Integrating the analysis of quantitative and qualitative data in one single study allowed it to answer the research questions proposed, and provided more in-depth findings. The mixed methods were employed in an *embedded design*. In this design, quantitative, and qualitative data are used to answer different research questions within a study (Hanson et al. 2005). Thus, while the study relied on the analysis of a quantitative dataset to estimate the amount of problematic content on online platforms during election campaigns, the interviews served to explore how experts evaluate the risks this problematic content may pose to elections and how to mitigate them.

1.2.2. Case selection

A case study is a typically qualitative research method to investigate "a contemporary phenomenon (the "case") in depth and within its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident....[and] relies on multiple sources of evidence" (Yin 2018, 45), allowing the study of complex social phenomena. Moreover, the case study approach allows in-depth description and multi-faceted explorations, while also analysing, comparing, and understanding different aspects of a research problem in its natural context (Crowe et al. 2011). Therefore, it is an appropriate empirical method when researchers need to gain contextual, concrete, and in-depth knowledge of complex issues in their real-life settings.

As part of the research design, this study used two case studies to assess the systemic risks online platforms pose to democratic elections.

`

Therefore, we conducted a ‘collective case study’, which involves studying multiple cases simultaneously or sequentially to generate a broader appreciation of a particular issue (Stake 1995). “In collective or multiple case studies, a number of cases are carefully selected. This offers the advantage of allowing comparisons to be made across several cases and/ or replication. Choosing a “typical” case may enable the findings to be generalised to theory (i.e. analytical generalisation) or to test theory by replicating the findings in a second or even a third case (i.e. replication logic)” (Crowe et al. 2011, 6).

Facebook and Twitter were chosen because they are the most relevant VLOPs globally, as well as in Germany. Thus, we believe that it is possible to better understand how an electoral risk assessment could or should be done by studying these specific platforms.

1.2.3. Data collection

This study analysed the three systemic electoral risk categories created (dissemination of illegal content, negative effects on electoral rights, and disinformation) on two global VLOPs operating in Germany: Facebook and Twitter. The assessment was based on representative samples of public data. This approach should enable us to make more reliable statements, allowing for meaningful comparisons of online platforms rather than the anecdotal data that is mostly used at present. The data collection period covered the second half of May 2021, from day 15 to day 31.

Every social media platform is organised in a slightly different way. For instance, on Twitter, hashtags are used to specify the topic or intended audience of a tweet and allow a user to engage a much larger potential audience than only his or her immediate followers. Hence, data to study the public debate of elections on Twitter can be collected through the use of one or more relevant hashtags and the subsequent analysis of the resulting universe of messages (Bruns and Burgess 2011; Larsson and Moe 2012; Lin et al. 2014; Shamma, Kennedy, and Churchill 2009).

`

Hashtags serve here as an indicator that a user's messages contributed to a given topic.[4] Facebook also allows the use of hashtags; however, it is not a primary feature of the platform, and they are not as routinely used as on Twitter.

To achieve a comparable sample of posts on both platforms, we collected data combining keywords and hashtags. Given that datasets collected using keywords risk introduce noise from the large number of messages using the keywords without actually referring to the intended topic(s), we chose keywords that referred uniquely to the election at hand, namely "Bundestagswahl" (federal election) and the abbreviations "BTW2021" and "BTW21". Given that the data was collected relatively early into the campaign devoid of major mediated events, we included hashtags of all political parties—Christian Democratic Union of Germany (CDU), Christian Social Union in Bavaria (CSU), Free Democratic Party (FDP), Grüne, Linkspartei, Social Democratic Party of Germany (SPD), and Alternative for Germany (AfD)—with a realistic chance of passing the electoral threshold of 5% required for representation in the Bundestag (Stier et al. 2018, 57). Furthermore, we included the names of each party's lead candidate ("Spitzenkandidat"), and hashtags already in use to refer to the election (#btw2021 and #btw21), as well as more general terms, such as #bundestagswahl and #wahlkampf.[5]

Twitter hashtags					
CDU	CSU	SPD	FDP	Gruene/ grüne	AfD
Linke	Linkspar tei	Bundestagswa hl	BTW202 1	BTW21	Wahlkam pf

[4] While an individual can engage in political communication without including a hashtag, the potential audience for such content is limited primarily to his or her immediate followers.

[5] In previous research, some researchers choose to collect data on what emerges to be the commonly used hashtag indicating content relevant to the upcoming election, such as #val2010 (Swedish for #election2010) (Larsson and Moe 2012) or #ausvotes (Bruns and Burgess 2011). Other studies cast a wider net, incorporating several hashtag linked to mediated events, such as the hashtags "tvduell" (referring to the broadcast of the debate between the two leading candidates), "petition", and "zensursula" (established during a campaign to support an e-petition against the law on access restrictions proposed by then Family Minister Ursula von der Leyen), in the dataset on the 2010 German election campaigns (Jürgens and Jungherr 2011, 6).

`

Chrupalla	Laschet	Scholz	Baerbock	Wissler	Bartsch
Lindner	Weidel				

On Twitter, we collected tweets from the platform's application programming interface (API) using the Python script *twarc2* and the Search endpoint. Our search returned 358.667 tweets in the period between 15 to 31 May 2021. Our Facebook dataset contained 6712 posts and 38.685 comments for the same time period.

Based on our previous research we estimated a response distribution of between 2% - 3%. We, thus, estimated that with a sample size of 1101 we would be able to able to attain a margin of error of 1% or lower with a 95% confidence interval. As a result, both datasets were subsequently transformed into randomised samples of 1101 entries each, and finally coded using the codebook previously created.

The resulting coding presented relatively high rates of intercoder reliability. For Facebook, with the coders agreeing that a post was problematic 93.10% of the time. For Twitter, the coders agreed that a post was problematic 92.55% of the time. These rates of intercoder reliability are within a good range that suggests reliable coding (McHugh 2012).

Finally, it is important to mention that during the data collection phase, several difficulties arose in accessing data from platforms. This situation makes it unnecessarily difficult, and sometimes impossible, for researchers to access reliable data to conduct research, undermining the capacity of third-party auditing of what happens on platforms and how effectively platforms enforce their policies.

`

2. Overview of the cases

The following sections briefly present Twitter and Facebook as the selected cases for our study. For this purpose, the history, basic functions, and characteristics of these platforms are explained. Furthermore, their strategies of content moderation and their approach regarding the management of potentially harmful content are highlighted, as well the role of advertisements in elections and political campaigns.

2.1. Twitter

Twitter is a micro-blogging platform on which users post and interact with short messages known as 'tweets'.[6] A profile is necessary to post, like, and retweet tweets; however, Twitter remains a relatively open platform, as tweets are accessible to read for the wider public without the need to register for an account. Short labels preceded by a hash mark, so-called #hashtags, which users freely create and use, serve to group tweets into conversations. Hashtags are the primary organisational feature of the platform, serving to categorise tweets according to what they are about (Moulaison and Burns 2012).

Founded in 2006, Twitter today serves an estimated 353 million active users a month worldwide (Hootsuite 2021). Its business model relies on the sale of clearly labelled, so-called promoted tweets, which are otherwise ordinary tweets purchased by advertisers, allowing them to reach a wider group of users. In 2019, Twitter announced that it would ban all political advertisements, defined as those sponsored by candidates or that discussed political issues, elections, candidates, parties, and overtly political content (Conger 2019). Twitter's audience in Germany was growing by 30% in the fourth quarter of 2020, which is the largest increase in all Twitter markets worldwide. According to the data made available to advertisers by Twitter, roughly 5.45 million Germans use this social media platform (Hootsuite 2020).

[6] In 2017, the original length of tweets comprising 140 characters was doubled to 280.

`

At its origins, Twitter practiced a low interference approach regarding content moderation, with its executive Tony Wang referring to the platform as “the free speech wing of the free speech party” (Halliday 2012) in 2012**.** The public concern about the circulation of disputed or misleading information related to COVID-19 prompted the company to introduce new features in May 2020, expanding the labels for tweets containing “synthetic and manipulated media” that had been “significantly and deceptively altered or fabricated” introduced earlier that year (Shapiro and Juhasz 2020). These new labels warn readers about potential misinformation and disinformation and provide additional context through links to trusted news resources (Y. Roth and Pickles 2020). Previously, tweet removal was the only tool used to enforce the platform’s ToS and legal requirements.

Twitter’s users are bound to observe its ‘rules’, which are a combination of legal requirements that Twitter is obliged to enforce by law and its ToS. The rules focus on three key areas: safety, privacy, and authenticity. The most extensive set of rules concern safety. Thus, under its safety rules, users are prohibited from engaging in the targeted harassment of other users, or inciting other people to do so, and from glorifying violence, terrorism, violent extremism, or threatening violence against an individual or a group. Furthermore, Twitter bans messages promoting suicide or child sexual exploitation, as well as graphic violence and adult content within live videos or in profile or header images. Additionally, Twitter bans the sale, purchase, or facilitation of transactions of illegal or certain types of regulated goods or services. Lastly, the platform bans hateful conduct defined as the promotion of violence, threats, or harassment of other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender identity, religious affiliation, age, disability, or serious disease.

Under the privacy layer, the company prohibits its users from publishing or posting other people’s private information without their authorisation or threatening to do so (e.g., so-called doxxing). It also bans the posting of

intimate photos or videos of someone who was produced or distributed without their consent.

With regards to ‘authenticity’, Twitter bars its users from violating others’ intellectual property rights, including copyright and trademark, and to impersonate individuals, groups, or organisations to mislead, confuse, or deceive others, or to share synthetic or manipulated media that are likely to cause harm. Platform manipulation and spam are outlawed. The rules explicitly state that the service may not be used “for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process” (Twitter n.d.).

In terms of rule enforcement, Twitter reports that in the second half of 2020, it removed 4.5 million unique pieces of content, such as tweets or an account’s profile image, banner, or bio (up 132% from the first half of 2020), and suspended 1 million accounts (up 9%) for violating the rules. Among the accounts permanently suspended was that of US president Donald Trump in early 2021, which generated a lot of controversy “due to the risk of further incitement of violence” regarding the criminal acts that took place in the U.S. Capitol on January 6 of that same year (Twitter 2021). A total of 3.5 million accounts were suspended or had some content removed (an 82% increase). The majority of content was removed for ‘hateful conduct’ (1.6 million, 77% increase) and ‘abuse/harassment’ (1.4 million, 142% increase). Twitter also reports that there was a 175% increase in content taken down for infringements of civic integrity, which was linked to the general U.S. elections in November 2020 (Twitter Transparency Center 2021).

2.2. Facebook

Facebook was launched in 2004 by Harvard psychology student Mark Zuckerberg using pictures from paper sheets distributed to college

`

freshmen. After 24 hours, more than 1000 students had signed up for the site (S. Phillips 2007). Zuckerberg had already created a social networking site called “Facemash,” which let users compare the attractiveness of Harvard students by using their photos from Harvard’s websites. The site was very successful among students but caused its founder some problems, as he was charged with breaching security, violating privacy, and copyrights (Carlson 2010). So-called student face books already existed at Harvard, but on thefacebook.com, which was the original name of the site, a collection of the entire student body’s pictures was created for the first time. Popularity increased, and the site was extended to other universities. By 2006, every person worldwide could sign up for the site as long as they were at least 13 years old (Barr 2018). Facebook reported 1.91 billion daily active users and 2.9 billion monthly active users, on average, in June 2021 (Facebook n.d.). In 2019, 32 million users per month and 23 million daily users were recorded in Germany (P. Roth 2019). Facebook has not published separate data about active German users for 2020 or 2021 (P. Roth 2021).

Every user who signs up to Facebook creates their profile, including pictures, personal information, and interests. This information can be shared publicly or only with a group of users. One of the main features of the platform is the so-called News Feed, in which users are shown status updates, posts, links, and pictures, likes, and reactions, and other activities happening on the platform. Users can share status updates, videos, and pictures, either publicly or with user groups (Facebook n.d.). The platform also enables private messaging to other users, video, and voice calls (Facebook n.d.). Users are required to use their names and provide correct information about themselves. They can only have one account for private purposes, which is not transferable. Moreover, children under 13 years of age, registered sex offenders, users who have been barred before, and users who are by law not allowed to use the service cannot register on the platform. Users are blocked, banned, or removed if they violate community standards or ToS (Facebook n.d.).

`

Facebook’s business model relies on allowing advertisers to place highly specific advertisements to narrowly targeted groups or audiences (micro-target). These ads, which take the form of images, videos, story ads, dynamic ads (advertising a product a user has seen on a product page or put in their shopping cart already), lead ads (using forms), or augmented reality ads, are placed on the users’ start page or in their private messenger app. Advertisers choose a specific objective, such as reaching a new audience, getting people to install an app, or making people watch the advertisers’ videos. Advertisers can select their target audience according to specific characteristics, such as location, age, gender, interests, and behaviours directly on the platform (Newberry 2020).

Facebook’s community standards prescribe the actions to be taken against various types of content. Thus, content that depicts violence, threats, and criminal behaviour, as well as any terrorist or violent organisations, illegal trades, or activities, are banned. Furthermore, content that includes suicide or self-harm, (sexual) exploitation, or personally identifiable information violating the privacy of individuals is prohibited and removed. Objectionable content, such as hate speech, depictions of violence, nudity, and sexual acts, offering commercial sexual services, or indelicate content, such as mockery, is prohibited. The community standards describe authenticity and integrity as crucial standards for the platform. Therefore, users are required to use their real names for their profiles. Accounts not complying with this standard are warned, and if violations are repeated, they are removed. Distributing spam or committing fraud is prohibited, and the use of fake accounts and identity disguises is also banned.

Moreover, disinformation and misinformation are countered by reducing economic incentives for people and pages distributing such content. Independent fact-checkers are consulted, and with their help, the distribution of flagged content is reduced. Manipulated media, such as deep fakes, are removed. Content violating community standards, such as spam, content that endangers the safety of people, such as sexual

`

exploitation, mobbing, or the violation of privacy, is removed. Moreover, content that includes nudity, hate speech and disinformation are regulated (Facebook n.d.). In the Community Standards Enforcement Report of the first quarter of 2021, Facebook reported having removed, among others, 1.3 billion fake accounts, 8.8 million pieces of content including bullying or harassment, and 25.2 million pieces of content comprising hate speech (Rosen 2021). Facebook prohibits posts that infringe on another's intellectual property rights, including copyrights, trademarks, and counterfeits. Users can report content that they think infringes their intellectual property (Facebook Transparency Center n.d.). However, according to Facebook, the majority of content infringing on intellectual property is removed proactively, accounting for 99.7% of all counterfeit removals and 77.9% of all copyright removals. Violations are identified using machine learning tools, information about prior violations, and common keywords used, especially for counterfeits (Fiore 2021). According to interviewed participants, however, information about the number of people exposed to pieces of content, including hate speech, and their frequency, would be much more valuable from a societal point of view. Reporting the prevalence, as Facebook currently does, does not further the understanding of the impact of the content on users, and it does not shed light on how the content came to be.

In October 2020, Facebook announced a ban on political advertisements on their platform, leading up to the US presidential elections. The ban was intended to limit the spread of disinformation concerning the elections (Paul 2020). The ban, however, has been called "much ado about nothing" (Kovach 2020), as it only included advertisements submitted after October 27, 2020. All other advertisements could still run, including targeted ads. Moreover, the measure has been criticised for being too late, as many people would already have voted. Spreading misinformation would not be affected, and political candidates could still spread misinformation about the results of elections (Kovach 2020). The ban was, however, lifted again in March 2021 (Schneider 2021). According to Facebook's ToS, advertisements for electoral campaigns are possible right now as long as

`

they have been authorised by the site and comply with legal regulations of the respective country. Furthermore, advertisements can still target specific audiences. An exception is the US state of Washington, where advertisements concerning elected officials, candidates, or election initiatives are not permitted (Facebook n.d.).

Facebook implemented an Oversight Board that decides on the content or revises decisions Facebook makes regarding content on its platform. The board was created in 2019 to enable an external independent appeals process by people not working for Facebook (Klonick 2020). The board understands itself as independent from Facebook, deciding about content and making policy recommendations. According to its charter, the board's decisions are binding on Facebook and need to be implemented (Oversight Board n.d.). The oversight board is financed by Facebook, which created a trust that pays the trustees overseeing the board. Therefore, the oversight board was criticised for not being independent. One of the most discussed decisions the board made was to prohibit former US President Donald Trump's account on the platform (O'Sullivan 2021).

`

3. Applying the DSA's risk assessment and mitigation framework to the German federal elections

The analysis of our representative samples of Facebook and Twitter data found a significant number of problematic posts and tweets. For the Facebook sample, 6.72% of all election-related posts were potentially illegal, disinformation, or infringements of electoral rights. As this data is based on the coding of a sample, it is possible that out sample overrepresents or underrepresents the underlying population data. However, with a confidence level of 95%, we can say that the underlying population data is within a margin of error of 1.46% of our sample. Of the problematic posts on Facebook, 4.05% were likely illegal under German law, 35.14% violated the platform's community standards or ToS, 46,65% were violations of electoral rights, and 93.24% could be considered disinformation.

Similarly, for the Twitter sample, 5.63% were found to be problematic. As these results are also based on the coding of a sample, here too our sample might be overrepresenting or underrepresenting certain categories within the overall population of Twitter data we analysed. However, we can say with a confidence level of 95%, that there is a margin of error of 1.34% between the Twitter sample we coded and overall population being studied. Of these problematic posts on Twitter, 14.52% broke platform rules, 51.61% infringed on electoral rights, and 100% were considered disinformation.

3.1. Dissemination of illegal content

Of the items flagged, 3 items (4.05%) in the Facebook sample and none in the Twitter sample were coded as likely illegal under German law. With regards to infringements of the service's ToS or community standards, the study identified 35.14% on the Facebook sample, and 14.52% on the Twitter sample.

`

Therefore, even after undergoing content moderation processes, there remained three potentially illegal posts on Facebook. We identified one post as 'malicious gossip (Üble Nachrede)' (subcategory code 2-1-17)[7] as described in §186 German Criminal Code, one post as 'disturbing public peace by threatening to commit offences' (2-1-7) as described in §126 German Criminal Code, and one post as 'incitement of masses' (Volksverhetzung, 2-1-11) based on the § 130 German Criminal Code that punishes incitement to hatred against segments of the population and refers to calls for violent or arbitrary measures against them. We did not find any illegal content in the Twitter sample.

In the course of our interviews, our experts pointed out that major risks are created through the absence of legal frameworks adapted to the online environment as well as a lack of enforcement, with governments granting platforms a great deal of latitude to decide what content should remain on platforms and what should be removed. Thus, a legal and human rights scholar highlighted the risk of the legal vacuum in which online platforms operate, as some statements would be illegal in traditional media but are merely considered harmful when made online. While content broadcast on television and published in print are subject to regulation, content on social media is largely not (Interviewee 4). This includes content that could encourage people to commit violent acts against some sections of society (Interviewee 1). To mitigate the risks from illegal content on online platforms, our interviewees stressed that these platforms are usually diligent when it comes to complying with rules; thus, adapting, and strengthening existing laws and enforcing them would help the platforms set an objective standard (Interviewee 1).

Our interviewees also emphasized that online platforms may have become the place where a wide variety of political tactics are being implemented, but the primary offenders are the political actors engaging in them, not the platforms as such. Thus, anyone who breaks the law should be held

[7] Hereafter, after the name of the subcategory, its code will be added in parentheses. See the 'codebook' developed for this study in the Annex section for more information on the subcategories and codes used.

`

accountable. “We need to consider the root causes of illegal content, and those are local participants in the election who are cheating. In elections, people cheat. If they do it on internet platforms, we have to prosecute them if it’s illegal,” said Interviewee 1. In a similar vein, another participant pointed out that authorities should not delegate the decision of what is legal and what is not to the platforms, as this is the function of a legal order.

Currently, online platforms exercise a great deal of discretion regarding the ban of accounts on their platforms, which includes those of politicians and political candidates. So far, it is unclear how political candidates would react when these platforms ban them and how courts might protect political candidates’ rights when they claim their freedom of expression was infringed upon. Several interviewees also mentioned that the platforms accommodate politicians who break rules to protect their business from retaliation. They mentioned the example of US President Donald J. Trump, whose accounts on Facebook and Twitter were only sanctioned in November 2020, when it became clear that he was unlikely to win the election (Interviewees 2 and 9). They also raised the risk posed by the lack of legislation for online ads, as opposed to TV advertisements, for instance.

Lastly, a hate speech and legal expert (Interviewee 5) proposed involving the users of online platforms in the process, increasing their sense of ownership by giving them a greater role in content moderation and thereby building a normative community so people behave differently online. They pointed to examples such as the Dangerous Speech Project,[8] which studied how volunteers respond collectively to hatred and dangerous speech online and thereby transpose norms common outside online spaces into those platforms. A former platform executive (Interviewee 1) cautioned against giving volunteers a greater role in content moderation, citing that from a freedom of expression standpoint, involving community moderators would most likely lead to more content

[8] See https://dangerousspeech.org

being taken down because they would remove what personally upsets them. Especially during election periods, involving volunteers in content moderation might not solve the problem of real or perceived bias, but it would shift it from platform administrators to individual users.

3.2. Negative effects on electoral rights

Of the problematic content found, 46,65% were violations of electoral rights on Facebook and 51.61% on Twitter. Within the electoral rights category, 'Candidates - Electoral campaign' (E-13) was the most common. On Facebook, subcategory E-13 accounted for 41.89% of all flagged posts. On Twitter, they accounted for all posts flagged as infringing electoral rights, that is, 51,61% of all problematic content. The prevalence of this category underlines that spreading disinformation by "Actors interested in harming/promoting certain candidates or parties or increasing social and political divisions in society spread misinformation on the private lives of candidates, or disinformation on political intentions, connections and activities of candidates and parties, or false allegations of violating campaign rules in order to defame candidates and parties, manipulate public opinion or influence voting behaviour" (Codebook), which is the most commonly employed strategy to harm certain candidates.

On Facebook, our coders furthermore registered the presence of 2.70% of posts coded under the subcategory 'Integrity – Electoral results' (E-19), defined as "Election losers and their supporters make undocumented claims on electoral fraud to justify electoral defeat, delegitimise democratic election, and encourage electoral protests" (Codebook). Also 1.35% were coded in the subcategory 'Integrity – Counting and notification' (E-17), which is defined as "Elections losers and their supporters make undocumented claims on lost ballot boxes, and non-counted votes, or the manipulation of vote counts and election protocols

`

etc. to justify electoral defeat, question electoral results and delegitimising elections, encouraging electoral protests." (Codebook).

Moreover, 1.35% were coded as subcategory 'Procedural – Vote count' (E-8), which is defined as "Actors interested in delegitimising the elections spread disinformation on procedures of the vote count to disturb the electoral process, confuse voters and to prevent (certain) voters from voting" (Codebook). Lastly, 1.35% were coded as subcategory 'Candidates – Election polls' (E-14), which is defined as "Actors interested in (de-)legitimising the elections or harming/promoting certain candidates or parties publish fictitious, false, or supportive election polls to (de-)mobilise voters and/or influence both voter turnout and voters' decisions." (Codebook).

Major risks to electoral rights identified by the interviewees include outdated electoral laws unfit for the online sphere and a lack of institutional oversight, as well as platforms playing favours with politicians, third-party interference, limited capacities of platforms to adequately respond to local specificities, and the very design of platform algorithms.

According to our interviewees, outdated regulations pose the most serious risk to elections. Although the laws' principles of current electoral laws are sensible and should be maintained, rules on expenditure, transparency, and third-party involvement need to be adapted to cover online advertisements and to legislate what is shown and who pays for certain spaces (Interviewees 1 and 4). As a first step, there is a need to define what constitutes an online political advertisement (Interviewee 10). Possible mitigation strategies include limiting the display of advertisements on online platforms, as proposed by Article 27 of the DSA.

Another strategy could be the creation of an official registry of online political ads modelled on the "Wahlwerberegister" in Germany, as well as oversight boards. In addition to a registry, when political ads are permissible, they should be strictly regulated, given that some parties,

`

such as the AfD, are using advertisements extensively outside election periods as well (Interviewee 5). Former platform employees pointed out the need for independent oversight agencies, such as independent election regulators, to carry out analysis, give clear guidance to platforms, and thereby instil confidence in the public about the processes.

However, at the moment, independent oversight agencies are underpowered and limited in scope, so the platforms are in charge of both intelligence gathering and enforcement (Interviewee 1). There were also calls for an audit of companies' processes for making integrity decisions and proper inspections of raw data through government agencies to fully grasp the situation before drafting regulations (Interviewee 3). For this, platforms would, of course, need to grant access to the most relevant information, which could trigger privacy issues (Interviewees 4 and 9). Most importantly, the election regulator or observatory would need to start working on these issues a year ahead of the election to adequately ring the alarm before the damage is already done (Interviewee 2).

Besides the necessary reform of the legal framework and institutional oversight, the interviewees highlighted the risk posed by the use of online platforms by politicians to bypass journalists and address their constituencies directly. Some actors, such as the AfD party, use this tactic extensively. This can be highly problematic because of the lack of regulation of these channels, and thus the political actors can spread their views unchecked by the so-called fourth power without having to face potentially uncomfortable questions from journalists (Interviewee 5). This is aggravated by the policies of online platforms, which explicitly do not conduct fact checking for politicians (Interviewee 3). High profile politicians especially benefit from exception to the rules because the platforms may suffer consequences from enforcing their rules on them. Thus, important public figures are subject to a process called 'cross check,' which exempts them from automated actions, according to a former Facebook employee (Interviewee 8). This in turn creates a risk of real-life harms and outrage leading to people being injured, as seen in the

`

inflammatory tweets published by former US President Donald Trump on Twitter (Interviewee 6). To mitigate this risk, platforms would need to enforce their own rules and apply them equally, including to politicians (Interviewee 2).

While fear of foreign interference in elections through online platforms has been a mainstay in the public debate in recent years, it is less important than disinformation spread by real people in a coordinated fashion (Interviewee 9). Even so, third-party interference is a major electoral risk, be it by foreign state actors or by commercial enterprises, such as Cambridge Analytica. This is a difficult issue that cannot be solved by domestic legislation alone but would need to be addressed in the area of media policy. There are also implications for the right to privacy, for example, through the non-consensual data collection by third parties during elections (e.g., Cambridge Analytica) (Interviewee 4).

Democracy, societal mores, and cultural norms do not function the same way in all countries, so risks will differ in different parts of the world (Interviewee 9). A former platform executive noted that while Facebook is concerned about the abuse of its platform for electoral purposes as a matter of reputation, the company only has limited capacities to adequately respond to local specificities. Facebook tends to transpose whatever worked in the United States to other countries, even though the local rules on political ads, for example, can be very different, simply because it lacks staff who really understand how elections work in a given country (Interviewee 1).

Scholars interviewed for this study underlined that so far, there is no scientific evidence that disinformation on election issues has an impact on human behaviour or that the use of social media by political parties has any impact on elections (Interviewees 4 and 8). The platforms' ad-financed algorithm recommender systems are designed to optimise engagement and increase the time users spend on the platforms, with the potential side effect of driving political polarisation. The primary aim of online platforms is to entertain, not to inform. Political campaigns serve this

`

purpose by creating competition for the attention of the users by using conflicting narratives, which drives user engagement, such as shares, likes, or comments (Interviewee 8). Possible solutions include relaxing certain parameters of the algorithms to decrease the amount of tailored information (Interviewee 8) or forcing platforms to forego algorithms altogether in favour of chronological feeds by default (Interviewee 9). Similar recommendations can be found in Article 27 of the DSA, which suggests that VLOPs adapt content moderation or recommender systems to mitigate systemic risks.

3.3. Disinformation

Disinformation is the most common form of problematic content found by the coders in both Facebook and Twitter samples. Of all content flagged as problematic, 93.24% we believe is disinformation on Facebook, and 100% we believe is disinformation on Twitter. Within the disinformation category, ‘trolling’ (D-9), defined as “the act of deliberately posting offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation” (Wardle et al. 2018), was by far the most prevalent in both datasets, with 47.30% of problematic Facebook posts and 57.68% of tweets we believe are rumours.

Other disinformation items found were ‘rumours’ (D-6), referring to “stories whose truthfulness is ambiguous or never confirmed (gossip, innuendo, unverified claims)” (Peterson and Gist 1951), with 31.08% of problematic posts coded as rumours on Facebook and 29.03% on Twitter. There were also 13.51% of Facebook posts and 14.52% of problematic tweets coded as ‘conspiracy theories’ (D-3), which are “Stories without factual base as there is no established baseline for truth. They usually explain important events as secret plots by government or powerful individuals” (Zannettou et al. 2019). In addition, 4.05% of Facebook posts and 4.84% of tweets were coded as ‘fabricated’ (D-1), defined as “Stories that completely lack any factual base, 100% false. The intention is to

`

deceive and cause harm" (Wardle and Derakshan 2017), which can be styled as news articles to make them appear legitimate.

Only on the Facebook sample the study found 5.41% of problematic posts were coded as 'pseudo-science' (D-10), which promotes "information that misrepresents real scientific studies with dubious or false claims." (Kapantai et al. 2021). A lower amount of content was coded as 'hoaxes' (D-4), which are relatively complex and large-scale fabrications presented as legitimate facts, intended to cause material loss or harm to the victim (Rubin, Chen, and Conroy 2015), with 1.35% of problematic Facebook posts. Further, 1.35% of Facebook post was coded as 'imposter' (D-2), which is defined in this study as genuine sources that are impersonated with false, made-up sources to support a false narrative. This can be very misleading, since the source or author is considered a great criterion for verifying credibility (Kapantai et al. 2021).

The greatest risks emanating from disinformation identified by our interviewees were the creation of pocket communities, the failure of common counterstrategies, notably of fact-checking, and platforms attracting bad actors for political or financial gain. In fact, the very design of social media platforms and the algorithms used to feed users' information enable the creation of small, tightly connected online communities, whose highly involved members perceive a false sense of consensus of their views (Interviewee 3). Their feeling that a lot of people agree with an idea they tend to agree with is a direct product of the algorithm designed to keep them spending time on the platform.

Our interviewees confirmed that there are numerous challenges with fact checking, which is at the heart of the major online platforms' strategies against disinformation. A former Facebook employee highlighted that the process of fact-checking is time-consuming, and there is a great latency because content is usually only fact-checked when it is already circulating widely. This means that the harm has already been done, with the risk that consumers of disinformation distance themselves further from mainstream reality (Interviewee 9). An academic scholar added that fact

`

checking only works for those already inclined to it, while for those hostile to debunking, the strategy will backfire because it reinforces their impression that someone seeks to influence them or that they are unfairly censored (Interviewee 8). This idea is supported by internal research conducted by social media platforms, which shows that disinformation significantly attracts more engagement when fact checked to be incorrect (Interviewee 2).

One way to improve fact checking would be to invest more into co-operations with trusted flaggers (DSA, Article 27), but most importantly to raise awareness and increase digital literacy, for example by teaching the public the different mechanisms of biases, such as confirmation bias and echo chambers (Interviewee 8). It would also be sensible to reinforce the internal processes of online platforms to detect systemic risk (DSA, Article 27). Furthermore, a former platform employee stressed the need for data about who spreads the information and their goals and strategies, which could be collected in the form of a library (Interviewee 3).

Another risk represents online platforms creating incentives for bad actors who create billions of fake accounts and networks of pages and coordinate posts with similar content in a similar time window, targeting people with a certain political leaning with increasingly radical content (Interviewee 7). For instance, a network of pages in France targeted users who do not like President Macron and fed them anti-immigration, anti-Muslimism, and other hate-related content (Interviewee 2). Further, the possibility of reaching millions of people through social media creates business opportunities for the promotion of disinformation as a commercial activity. Elections then serve as an opportunity to create fabricated news for financial gain (Interviewee 1). When this content is well crafted, it risks influencing voters' decisions on who to vote for or their trust in the integrity of electoral processes, especially when it matches existing political narratives (Interviewee 9).

Strategies to decrease the amount of disinformation proposed by the interviewees included fighting sock puppets by increasing the costs of

creating accounts, for example, by allowing only accounts older than three months to create groups (Interviewee 3). Given that oftentimes users share content after reading the headline only, another strategy is to increase friction by making content sharing more cumbersome by asking users if they really want to share a given item (Interviewees 1 and 9).

Lastly, a major problem is the non-transparency with which the platforms deal with disinformation by restricting access to data for external researchers. Twitter is slightly more open to researchers, which means that models for bot detection, as with most research on disinformation, are solely based on Twitter data (Interviewee 7). The lack of available data results in flawed science built on a great deal of speculation.

`

4. Policy recommendations

Our research shows that there is far too much problematic content on platforms. Finding 6.72% of problematic content on Facebook and 5.63% on Twitter was far higher than we expected. In light of the widespread public debate about election-related challenges, we were not expecting this level of problematic content to still be present on platforms, even after platforms' content moderation and design interventions were being implemented and at such an early moment in the election campaign. As the elections heat up before 26 September 2021, the proportion of problematic content is likely to be even higher than what we found in May 2021. Consequently, we have developed the following policy recommendations:

1. **Platforms need to create more effective and sustainable response mechanisms to do more to safeguard elections.** Our research suggests that all existing measures are currently very far from being good enough. Neither Facebook nor Twitter is doing enough to remedy the current situation.

2. **Platforms should implement research-based recommendations to improve their mitigation measures to problematic content before and during elections.** Our research suggests that platforms are not sufficiently considering a large body of knowledge and research on how to mitigate risks to free and fair elections and democracy. This includes interface-design solutions and tools that can empower users in the online ecosystem. The platforms should thus be required under Art. 27 DSA to develop mitigation measures together with civil society organisations and independent experts. Criteria for cooperation should be defined where appropriate.

3. **Platforms have to become more transparent about content moderation tools they deploy, including algorithmic transparency.** In this vein, platforms should publicly disclose the number of false positives and false negatives, and what content is flagged by algorithms and trusted flaggers (so-called precision and recall data). Especially in the context of disinformation, the time and intensity of exposure combined with the visibility of disinformation content on platforms is meaningful information to better understand its spread online. Moreover, platforms should provide information on the extent to which they profit (intentionally or unintentionally) from systemic risks (e.g., estimations of turnover generated through disinformation or illegal content). These additional transparency requirements could be included under Art. 23 DSA.

4. **Platform terms of service need to be expanded to more effectively cover all forms of disinformation and electoral rights, especially in times of elections.** Only a small part of all the problematic content we found was covered by the existing terms of service of Twitter and Facebook.

5. **There is a need to for platforms to adopt best practices in their responses mechanisms to problematic content.** The differences between Twitter and Facebook suggest real differences in the quality of their responses to problematic content about elections. If Twitter is doing better than Facebook despite Facebook's dominant economic position in the market, Facebook should be expected to do at least as well as Twitter. However, both could still do significantly better.

6. **Almost all the problematic content we found was legal content.** Platforms should be obliged to disclose how they distinguish between permissive and illegal content and conduct risk assessments for the types of legal but problematic content we discuss in this report. Public disclosure of such information may

decrease uncertainty among users and, at the same time, increase the trust in platforms' content moderation processes. Furthermore, we suggest, in line with our previous research (Tiedeke et al. 2020; Wagner et al. 2021), that just focusing on illegal content to safeguard elections will be ineffective.

7. **Platforms should focus on curation, moderation, and design measures that promote free expression and user agency over the information they receive and impart.** In their risk mitigation measures to safeguard elections, platforms should focus primarily on design changes and other measures more likely to promote free expression. Content moderation is clearly also necessary but is more likely to cause problems and needs to be done in a transparent and accountable manner.

8. **Categories of analysis need to be improved.** Despite a relatively high intercoder reliability rate, we often struggled to clearly identify the boundaries of the categories for identifying electoral rights violations and disinformation. Comparatively, the legal categories were easier to operationalise and more clearly delineated. Our interviewees also suggested that the existing categories of systemic electoral risk in the DSA and relevant academic literature still need to be more clearly delineated and easier to reproduce and compare. Even though our categories are based on the DSA and state-of-the-art academic literature, we believe that additional research and policy development is needed to operationalise and clearly delineate what constitutes disinformation and electoral rights violations.

9. **EU legislators should expand DSA risk assessments and DSA Article 29 transparency criteria beyond very large online platforms.** Our research has focused primarily on large online platforms. However, we agree with many of the experts we interviewed, who suggested that smaller platforms can also have highly problematic effects on elections (EU DisinfoLab 2020,

Shalvey 2021). Particularly, given that the boundary between very large online platforms and other platforms in the DSA (10% of all EU citizens) seems highly arbitrary, we suggest implementing relevant parts of the risk-based approach beyond only VLOPs alone.

To do this effectively, regulators need to focus – beyond those platforms that are already covered by the DSA and should remain so- on the impact platforms can have rather than the number of users they have. In the context of elections, this means that all platforms where there is scientific evidence that the platform can influence an EU Member State's election or an EU election should be included as part of the risk-based approach. Scientific studies of this kind already exist for some platforms (Bond 2012), but further impact assessments would obviously be needed for other platforms as well. Furthermore, these rules should also apply to video-sharing platforms.

10. **Researchers need better access to platform data.** Access to data remains highly challenging and politically charged. It was very difficult and time consuming to gain access to the representative samples we needed to conduct this research. After gathering the data, we constantly felt concerned about arbitrary risks to ourselves and our partners. The experience of New York University being shut out by Facebook based on claims of privacy violations (Vincent 2021) and an almost universal fear among the community conducting this research creates legal conflicts with the platforms. This is no way to conduct research, as it has a chilling effect on the ability of researchers to hold platforms accountable. Under Art. 31 DSA, vetted researchers must thus be granted access to relevant data. This is to enable research that contributes to a better understanding of systemic risks as well as of the underlying economic incentives for platforms on how to deal with them.

11. **Auditors must be chosen and paid by the authorities.** The DSA relies heavily on independent auditors to examine systemic electoral risks and develop effective mitigation measures. The present external risk assessment-despite being a smaller version-might probably not have been commissioned by a platform due to its critique. However, public auditing intermediaries should be introduced to further secure and strengthen the independence of auditors and the auditing regime (Wagner and Kuklis 2021). Finally, to ensure auditors' independence, it is crucial to clarify under Art. 28 DSA that auditors are commissioned by the envisaged European Board for Digital Services.

`

5. Final considerations

Safeguarding democratic elections is hard. We acknowledge that online platforms and their regulators have an enormously difficult task ahead of them in trying to safeguard elections. However, this acknowledgement should not detract from the fact that private platforms and public regulators' current efforts to safeguard elections are simply not sufficient. As a result, democratic elections will continue to suffer from disinformation and continuous breaches of electoral rights.

Social media platforms are not a mirror of society, even if they often like to claim so. Their presence in society has effects that cannot be taken for granted, nor are they likely to go away any time soon. Regulators need to acknowledge the central role of these platforms in elections and systematically develop institutions that are adequately able to respond to the issues discussed. These institutions urgently need to be strengthened both in Germany and at the EU level.

The EU DSA can undoubtedly contribute to improving the mitigation of systemic risks from the platforms. In particular, Article 26 and Article 27 of the DSA studied here create a valuable regulatory framework to push these platforms in the right direction. However, without expanded external audits of the platforms, they will continue to run rings around regulators and election observers. "[T]hey're playing us" (Wagner 2020, 743), one leading election observer acknowledged, even as he spent his days "running after the tech companies." (Wagner 2020, 743).

Despite these regulatory developments, public institutions often fail to compel platforms to do better. The more politically independent the electoral authority and the more tech-savvy their staff, the more likely they are to push the platforms in their right direction. Voters can and should demand better public institutions as well as better private platforms.

`

Importantly, the idea frequently stated by current and former Facebook staff that elections are ‘on balance’ better than they previously were before social media lacks empirical foundation. We don’t know what democratic elections would look like without social media. Still, we can legitimately claim that elections would not be democratic elections if social media were not present or completely censored. The relevant question is not whether democratic elections are compatible with social media but rather how online platforms can be developed further to be more supportive of free and fair elections. This will likely require considerable resources and probably take some time, but it is definitely not impossible. If anything, it seems that these platforms are not sufficiently considering the vast body of knowledge that already exists, and even some of their internal research (Hao 2021). If this is not taken seriously, safeguarding democratic elections is essentially impossible.

However, it does not have to be this way. We know that different performances by online platforms are possible by comparing how well the large platforms perform and are even more possible by considering many of the smaller online platforms that do a better job. The question is whether platforms and their regulators will be willing to take the systemic risks around elections seriously and take meaningful steps to mitigate them. These platforms should not simply be doing this a little here and there before each election campaign but instead systematically building more sustainable platforms.

`

References

Anand, Bharat N. 2017. ‘The U.S. Media’s Problems Are Much Bigger than Fake News and Filter Bubbles’. *Harvard Business Review*, 5 January 2017. https://hbr.org/2017/01/the-u-s-medias-problems-are-much-bigger-than-fake-news-and-filter-bubbles.

Barr, Sabrina. 2018. ‘When Did Facebook Start? The Story Behind a Company That Took over the World.’ *The Independent*, 23 August 2018, sec. Lifestyle. https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-when-started-how-mark-zuckerberg-history-harvard-eduardo-saverin-a8505151.html.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam D I Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. ‘A 61-Million-Person Experiment in Social Influence and Political Mobilization’. *Nature* 489 (7415): 295–98.

Bovet, Alexandre, and Hernán A. Makse. 2019. ‘Influence of Fake News in Twitter During the 2016 US Presidential Election’. *Nature Communications* 10 (1): 7. https://doi.org/10.1038/s41467-018-07761-2.

Bruns, Axel, and Jean Burgess. 2011. ‘New Methodologies for Researching News Discussion on Twitter’. In *Proceedings of the 3rd Future of Journalism Conference*, edited by A. Phillips, 1–11.

Carlson, Nicholas. 2010. ‘At Last — the Full Story of How Facebook Was Founded’. *Business Insider*, 5 March 2010. https://www.businessinsider.com/how-facebook-was-founded-2010-3.

Charter of Fundamental Rights of the European Union. 2012. *2012/c 326/02*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN.

Conger, Kate. 2019. ‘What Ads Are Political? Twitter Struggles With a Definition’. *The New York Times*, 15 November 2019, sec. Technology. https://www.nytimes.com/2019/11/15/technology/twitter-political-ad-policy.html.

Crowe, Sarah, Kathrin Cresswell, Ann Robertson, Guro Huby, Anthony Avery, and Aziz Sheikh. 2011. ‘The Case Study Approach’. *BMC Medical Research Methodology* 11 (June): 100. https://doi.org/10.1186/1471-2288-11-100.

EU DisinfoLab. 2021. ‘How the Digital Services Act (DSA) Can Tackle Disinformation’. EU DisinfoLab. https://www.disinfo.eu/advocacy/how-the-digital-services-act-(dsa)-can-tackle-disinformation/.

European Commission. 2018a. ‘Free and Fair European Elections’.

———. 2018b. ‘Securing Free and Fair European Elections’. COM(2018) 637 final. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0637&from=EN.

———. 2020. *Proposal for Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC*. *2020/0361 (COD)*. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM

%3A2020%3A825%3AFIN.
———. 2021. ‘European Commission Guidance on Strengthening the Code of Practice on Disinformation’. May 2021. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.
Facebook. n.d. ‘Facebook Community Standards’, Section IV’. https://www.facebook.com/communitystandards/integrity_authenticity/.
———. n.d. ‘Facebook Nutzungsbedingungen’. Accessed 9 August 2021a. https://www.facebook.com/terms.php.
———. n.d. ‘Facebook Reports Second Quarter 2021 Results’. Accessed 9 August 2021b. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Second-Quarter-2021-Results/default.aspx.
———. n.d. ‘Gemeinschaftsstandards’. Accessed 9 August 2021c. https://www.facebook.com/communitystandards/introduction.
———. n.d. ‘How News Feed Works | Facebook Help Center’. Accessed 9 August 2021d. https://www.facebook.com/help/1155510281178725/?locale=en_US.
———. n.d. ‘Werberichtlinien’. Accessed 9 August 2021e. https://www.facebook.com/policies/ads/restricted_content/political#.
Facebook Transparency Center. n.d. ‘Intellectual Property’. Accessed 9 August 2021. https://transparency.fb.com/data/intellectual-property.
Fiore, Mark. 2021. ‘How We’re Proactively Combating Counterfeits and Piracy’. *About Facebook* (blog). 19 May 2021. https://about.fb.com/news/2021/05/how-were-proactively-combating-counterfeits-and-piracy/.
Halliday, Josh. 2012. ‘Twitter’s Tony Wang: “We Are the Free Speech Wing of the Free Speech Party”‘. The Guardian. 22 March 2012. http://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.
Hanson, William E., John W. Creswell, Vicki L. Plano Clark, Kelly S. Petska, and J. David Creswell. 2005. ‘Mixed Methods Research Designs in Counseling Psychology.’ *Journal of Counseling Psychology* 52 (2): 224–35. https://doi.org/10.1037/0022-0167.52.2.224.
Hao, Karen. 2021. ‘How Facebook Got Addicted to Spreading Misinformation’. MIT Technology Review. 2021. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.
HLG. 2018. ‘A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation’. LU: Publications Office. https://data.europa.eu/doi/10.2759/739290.
Hootsuite. 2020. ‘Digital 2020: October Update’. 2020. https://www.hootsuite.com/resources/digital2020-q4-update.
———. 2021. ‘36 Twitter Statistics All Marketers Should Know in 2021’. *Social Media Marketing & Management Dashboard* (blog). 3 February 2021. https://blog.hootsuite.com/twitter-statistics/.
Howard, Philip N., and Bence Kollanyi. 2016. ‘Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum’. SSRN Scholarly Paper ID 2798311. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.2798311.

`

Inter-Parliamentary Council. 1994. *Declaration on Criteria for Free and Fair Elections*. https://www.ipu.org/our-impact/strong-parliaments/setting-standards/declaration-criteria-free-and-fair-elections.

Ireton, Cherilyn, Julie Posetti, and UNESCO. 2018. *Journalism, 'Fake News' et Disinformation: Handbook for Journalism Education and Training*. http://unesdoc.unesco.org/images/0026/002655/265552E.pdf.

Isaac, Mike. 2016. 'Facebook, in Cross Hairs After Election, Is Said to Question Its Influence'. *The New York Times*, 12 November 2016, sec. Technology. https://www.nytimes.com/2016/11/14/technology/facebook-is-said-to-question-its-influence-in-election.html.

Juncker, Jean-Claude. 2018. 'State of the Union 2018: European Commission Proposes Measures for Securing Free and Fair European Elections'. Text. European Commission - European Commission. 2018. https://ec.europa.eu/commission/presscorner/detail/en/IP_18_5681.

Jürgens, Pascal, and Andreas Jungherr. 2011. 'Wahlkampf vom Sofa aus: Twitter im Bundestagswahlkampf 2009'. In *Das Internet im Wahlkampf*, edited by Eva Johanna Schweitzer and Steffen Albrecht, 201–25. Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92853-1_8.

Kapantai, Eleni, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. 'A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework'. *New Media & Society* 23 (5): 1301–26. https://doi.org/10.1177/1461444820959296.

Karlova, Natascha A., and Karen E. Fisher. 2013. 'A Social Diffusion Model of Misinformation and Disinformation for Understanding Human Information Behaviour'. *Information Research* 18 (1).

Klonick, Kate. 2020. 'The Facebook Oversight Board: Creating an Independent Institution to Adjucate Online Free Expression'. *Yale Law Journal* 129 (8): 2422–99.

Kovach, Steve. 2020. 'Facebook's Ban on New Political Ads Won't Change Anything'. *CNBC*, 3 September 2020, sec. Technology. https://www.cnbc.com/2020/09/03/facebooks-ban-on-new-political-ads-wont-change-anything.html.

Lapowsky, Issie. 2016. 'This Is How Facebook Actually Won Trump the Presidency'. *Wired*, 15 November 2016. https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/.

Larsson, Anders Olof, and Hallvard Moe. 2012. 'Studying Political Microblogging: Twitter Users in the 2010 Swedish Election Campaign'. *New Media & Society* 14 (5): 729–47. https://doi.org/10.1177/1461444811422894.

Leetaru, Kalev. 2019. 'The Importance of Context and Intent in Content Moderation'. Forbes. 28 July 2019. https://www.forbes.com/sites/kalevleetaru/2019/07/28/the-importance-of-context-and-intent-in-content-moderation/.

Lin, Yu-Ru, Brian Keegan, Drew Margolin, and David Lazer. 2014. 'Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events'. Edited by Petter Holme. *PLoS ONE* 9 (5): e94093. https://doi.org/10.1371/journal.pone.0094093.

Manheim, David. 2020. 'The Fragile World Hypothesis: Complexity,

Fragility, and Systemic Existential Risk'. *Futures* 122 (September): 102570. https://doi.org/10.1016/j.futures.2020.102570.
Matsa, Katerina Eva. 2018. 'Most Western Europeans Get News from TV as Print Reading Lags'. *Pew Research Center* (blog). 27 September 2018. https://www.pewresearch.org/fact-tank/2018/09/27/most-western-europeans-prefer-tv-news-while-use-of-print-outlets-lags/.
McHugh, Mary L. 2012. 'Interrater Reliability: The Kappa Statistic'. *Biochemia Medica* 22 (3): 276–82.
Mihas, Paul and Odum Institute. 2019. *Learn to Build a Codebook for a Generic Qualitative Study*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications, Ltd. https://doi.org/10.4135/9781526496058.
Moulaison, Heather Lea, and C. Sean Burns. 2012. 'Organization or Conversation in Twitter: A Case Study of Chatterboxing'. *Proceedings of the American Society for Information Science and Technology* 49 (1): 1–11. https://doi.org/10.1002/meet.14504901185.
Newberry, Christina. 2020. 'How to Advertise on Facebook: Complete Facebook Ads Guide for 2021'. *Social Media Marketing & Management Dashboard* (blog). 9 March 2020. https://blog.hootsuite.com/how-to-advertise-on-facebook/.
O'Sullivan, Donie. 2021. 'What You Need to Know about the Board Deciding Trump's Fate on Facebook'. *CNN Business*, 4 May 2021. https://edition.cnn.com/2021/05/04/tech/what-is-facebook-oversight-board/index.html.
Oversight Board. n.d. 'Governance'. Accessed 9 August 2021. https://oversightboard.com/governance/.
Paul, Kari. 2020. 'Facebook Announces Plan to Stop Political Ads after 3 November'. *The Guardian*, 8 October 2020, sec. Technology. http://www.theguardian.com/technology/2020/oct/07/facebook-stop-political-ads-policy-3-november.
Peterson, Warren A., and Noel P. Gist. 1951. 'Rumor and Public Opinion'. *American Journal of Sociology* 57 (2): 159–67. https://doi.org/10.1086/220916.
Phillips, Sarah. 2007. 'A Brief History of Facebook'. *The Guardian*, 25 July 2007, sec. Technology. http://www.theguardian.com/technology/2007/jul/25/media.newmedia.
Rhodes, Samuel C. 2021. 'Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation'. *Political Communication* 0 (0): 1–22. https://doi.org/10.1080/10584609.2021.1910887.
Roberts, Sarah T. 2017. 'Content Moderation'. In *Encyclopedia of Big Data*, edited by Laurie A. Schintler and Connie L. McNeely, 1–4. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_44-1.
Rogers, Richard. 2020. 'Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media'. *European Journal of Communication* 35 (3): 213–29. https://doi.org/10.1177/0267323120922066.
Rosen, Guy. 2021. 'Community Standards Enforcement Report, First

Quarter 2021’. *About Facebook* (blog). 19 May 2021. https://about.fb.com/news/2021/05/community-standards-enforcement-report-q1-2021/.

Roth, Philipp. 2019. ‘Offizielle Facebook Nutzerzahlen für Deutschland (Stand: März 2019)’. *allfacebook.de* (blog). 19 March 2019. https://allfacebook.de/zahlen_fakten/offiziell-facebook-nutzerzahlen-deutschland.

———. 2021. ‘Nutzerzahlen: Facebook, Instagram, Messenger und WhatsApp, Highlights, Umsätze, uvm. (Stand April 2021)’. *allfacebook.de* (blog). 29 April 2021. https://allfacebook.de/toll/state-of-facebook.

Roth, Yoel, and Nick Pickles. 2020. ‘Updating Our Approach to Misleading Information’. 11 May 2020. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

Rubin, Victoria L., Yimin Chen, and Nadia K. Conroy. 2015. ‘Deception Detection for News: Three Types of Fakes’. *Proceedings of the Association for Information Science and Technology* 52 (1): 1–4. https://doi.org/10.1002/pra2.2015.145052010083.

Shalvey, Kevin. 2021. ‘As Clubhouse’s popularity skyrockets, some observers are raising questions about the spread of misinformation’. Business Insider France. 20 March 2021. https://www.businessinsider.fr/us/clubhouses-growth-raises-troubling-questions-over-spread-of-misinformation-2021-3.

Schneider, Elena. 2021. ‘Facebook Lifts Political Ad Ban’. *POLITICO*, 3 March 2021. https://www.politico.com/news/2021/03/03/facebook-lifts-political-ad-ban-473368.

Shamma, David A., Lyndon Kennedy, and Elizabeth F. Churchill. 2009. ‘Tweet the Debates: Understanding Community Annotation of Uncollected Sources’. In *Proceedings of the First SIGMM Workshop on Social Media - WSM ‘09*, 3. Beijing, China: ACM Press. https://doi.org/10.1145/1631144.1631148.

Shapiro, Ari, and Aubri Juhasz. 2020. ‘Twitter Vows That As Disinformation Tactics Change, Its Policies Will Keep Pace’. NPR.Org. 4 March 2020. https://www.npr.org/2020/03/04/811686225/twitter-vows-that-as-disinformation-tactics-change-its-policies-will-keep-pace.

Spohr, Dominic. 2017. ‘Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure on Social Media’. *Business Information Review* 34 (3): 150–60. https://doi.org/10.1177/0266382117722446.

Stake, Dr Robert E. 1995. *The Art of Case Study Research*. Thousand Oaks: SAGE Publications, Inc.

Stier, Sebastian, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. 2018. ‘Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter’. *Political Communication* 35 (1): 50–74. https://doi.org/10.1080/10584609.2017.1334728.

Tiedeke, Anna-Sophia, Matthias C. Kettemann, Felicitas Rachinger, Marie-Therese Sekwenz, and Ben Wagner. 2020. ‘What Can Be Said Online in Germany and Austria? A Legal and Terms of Service Taxonomy’.

SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3735932.
Twitter. 2021. ‘Permanent Suspension of @realDonaldTrump’. 8 January 2021. https://blog.twitter.com/en_us/topics/company/2020/suspension.
———. n.d. ‘The Twitter Rules: Safety, Privacy, Authenticity, and More’. Accessed 23 July 2021. https://help.twitter.com/en/rules-and-policies/twitter-rules.
Twitter Transparency Center. 2021. ‘Rules Enforcement’. 14 July 2021. https://transparency.twitter.com/en/reports/rules-enforcement.html.
United Nations. 1948. ‘Universal Declaration of Human Rights’. United Nations. United Nations. 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.
Vincent, James. 2021. ‘Facebook Bans Academics Who Researched Ad Transparency and Misinformation on Facebook’. The Verge. 4 August 2021. https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in.
Wagner, Ben. 2020. ‘Digital Election Observation: Regulatory Challenges Around Legal Online Content’. *The Political Quarterly* 91 (4): 739–44. https://doi.org/10.1111/1467-923X.12903.
Wagner, Ben, and Marina Kubina. 2021. ‘Ergebnisse Des Forschungsprojekts Zur Stärkung Der Diskussionskultur - CommunityBlog - DerStandard.at › Diskurs’. 18 February 2021. https://www.derstandard.at/story/2000124046106/ergebnisse-des-forschungsprojekts-zur-staerkung-der-diskussionskultur.
Wagner, Ben, Johanne Kübler, Eliška Pírková, Rita Gsenger, and Carolina Ferro. 2021. ‘Reimagining Content Moderation and Safeguarding Fundamental Rights: A Study on Community-Led Platforms’. Tallinn, Estonia: Enabling Digital Rights and Governance & EU Greens/EFA. https://enabling-digital.eu/wp-content/uploads/2021/07/Alternative-content_web.pdf.
Wagner, Ben, and Lubos Kuklis. 2021. ‘Disinformation, Data Verification and Social Media: Verifying Data through Auditing Intermediaries’. In *Dealing with Digital Dominance*. Oxford, UK: Oxford University Press.
Wardle, Claire, and H. Derakshan. 2017. ‘Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making’. Council of Europe.
Wardle, Claire, Grace Greason, Joe Kerwin, and Nic Dias. 2018. ‘Information Disorder, Part 1: The Essential Glossary’. *First Draft* (blog). 2018. https://medium.com/1st-draft/information-disorder-part-1-the-essential-glossary-19953c544fe3.
Yin, Robert K. 2018. *Case Study Research: Design and Methods*. Sixth edition. Los Angeles, Calif: SAGE Publications Inc.
Zannettou, Savvas, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. ‘The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans’. *Journal of Data and Information Quality* 11 (3): 1–37. https://doi.org/10.1145/3309699.

`

6. Annexes

6.1. Codebook

The list of codes (codebook) used to code and analyse the Facebook and Twitter samples is expressed below, organised by category of systemic risks. A much bigger version of this codebook was developed for the study, with definitions and examples of each subcategory. However, due to its large extension, it was decided to present only the abridged version here.

6.1.1. Category: Dissemination of illegal content

1-1	Terrorism promotion
1-2	Terrorism financing
1-3	Sexual exploitation of children / Child abuse material or anything objectionable involving minors, grooming or predation
1-4	Promotion of and/or enabling of human trafficking, Offering or advertising human trafficking
1-5	Incitement to or promotion of genocide (jus cogens)
1-6	Qualified hate speech amounting to calls for violence
1-7	Qualified breaches of personal information, defamatory personal content, Impersonation (fake accounts/profiles/pages), sexual objectification, unauthorized dissemination of intimate images (“revenge porn”)
1-8	Qualified violations/misuses of freedom of expression
1-9	Prohibition of propaganda for war and inciting national, radical or religious hatred: Hate speech, violent extremist content.

2-1-1	Dissemination of propaganda material of unconstitutional organisations
2-1-2	Use of symbols of unconstitutional organisations
2-1-3	Preparation of serious violent offence endangering state
2-1-4	Instructions for committing serious violent offence endangering state
2-1-5	Treasonous forgery
2-1-6	Public incitement to commit offences
2-1-7	Disturbing public peace by threatening to commit offences. Expressions that are likely to disturb public peace.
2-1-8	Forming criminal organisations
2-1-9	Forming terrorist organisations
2-1-10	Foreign criminal and terrorist organisations
2-1-11	Incitement of masses (Volksverhetzung)
2-1-12	Depictions of violence
2-1-13	Rewarding and approval of offences

`

2-1-14	Revilement of religious faiths and religious and ideological communities
2-1-15	Dissemination, procurement and possession of child pornography, making pornographic content available through broadcasting or tele media services; accessing child or youth pornographic content via tele media
2-1-16	Insult
2-1-17	Malicious gossip (Üble Nachrede)
2-1-18	Defamation
2-1-19	Violation of intimate privacy by taking photographs or other images
2-1-20	Threatening commission of serious criminal offence
2-1-21	Forgery of data of probative value
2-1-22	Labelling and advertising / Criminal Offences
2-1-23	Advertising, obligations to inform, trade prohibitions Weapons Act / Criminal Offences
2-1-24	Drug Advertising Act /Criminal Offences

2-2-1	Disturbance liability (Störerhaftung)
2-2-2	Quasi-negatory injunctive relief and tort law
2-2-3	Copyright infringement
2-2-4	Trademark infringement
2-2-5	Contractual obligations
2-2-6	Spam

3-1	Violence and incitement
3-2	Content of dangerous individuals and organizations
3-3	Coordinated harm and publicizing crime
3-4	Content on regulated goods
3-5	Fraud and deception
3-6	Suicide and Self-Injury
3-7	Child Nudity and sexual exploitation of children
3-8	Sexual exploitation of adults
3-9	Bullying and Harassment
3-10	Human exploitation
3-11	Privacy violations and image privacy rights
3-12	Hate Speech
3-13	Violent Graphic Content
3-14	Adult Nudity and Sexual Activity
3-15	Sexual Solicitation
3-16	Cruel and insensitive content
3-17	Misrepresentation
3-18	Spam
3-19	Cybersecurity
3-20	Inauthentic Behaviour
3-21	False News
3-22	Manipulated Media
3-23	Memorialization

3-24	Intellectual Property
4-1	Does not meet the threshold of a legally relevant behaviour but can affect people's feelings of comfort. May lead to uncomfortable feelings.
4-2	Does not fit into a legally relevant category but can be considered to be equally relevant/invasive. Because of the intensity or extent. And because of the fact that there have not yet been cases decided by courts regarding these categories and new legislation has not been passed, yet.

6.1.2. Category: Negative effects on electoral rights

E-1	Procedural – Voter registration
E-2	Procedural – Right to vote (active suffrage)
E-3	Procedural – Voter identification
E-4	Procedural – Election campaign
E-5	Procedural – Election campaign and party funding
E-6	Procedural – Polling station
E-7	Procedural - Voting
E-8	Procedural – Vote count
E-9	Procedural – Election observation
E-10	Procedural – Electoral system
E-11	Candidates – Right to stand for elections
E-12	Candidates – Electoral registration of candidates and parties
E-13	Candidates – Electoral campaign
E-14	Candidates – Election polls
E-15	Integrity – voter registration
E-16	Integrity - Voting
E-17	Integrity – Counting and notification
E-18	Integrity – Publishing of electoral results
E-19	Integrity – Electoral results
E-20	Integrity – Electoral observation

6.1.3. Category: Disinformation

D-1	Fabricated
D-2	Imposter
D-3	Conspiracy theories
D-4	Hoaxes
D-5	Biased or one-sided *(not relevant for this study)*
D-6	Rumours
D-7	Clickbait
D-8	Misleading connection
D-9	Trolling
D-10	Pseudoscience
D-11	Fake reviews *(not relevant for this study)*

6.2. Abbreviations

AfD	Alternative for Germany
API	application programming interface
CDU	Christian Democratic Union of Germany
CSU	Christian Social Union in Bavaria
DSA	Digital Service Act
EU	European Union
FDP	Free Democratic Party
SPD	Social Democratic Party of Germany
ToS	terms of service
VLOPs	very large online platforms

6.3. Questionnaire

1. What is your role in relation to the DSA and online risk management?
2. How many years of experience do you have working in this field?
3. What do you think are the most prevalent risks of online platforms during elections?
4. How do you think these risks could influence human rights and free and fair elections during the election period?
5. How do you think these risks to free and fair elections and human rights should be assessed most accurately? And how should or could they be mitigated?
6. How do you think risks like online manipulation and disinformation could be prevented or mitigated?
7. Do you think existing content moderation practices and techniques by online platforms are effective at mitigating all of the risks we've discussed so far?
8. Large platforms like Twitter and Facebook have recently introduced design changes to inform users when they share information rated misleading by third-party fact-checkers. What role do you think the design of the platform itself plays in systemic risk mitigation?
9. What steps does/could a platform take to reduce systemic risks through the design of its platform?
10. We've developed some ideas of what design changes platforms could make to reduce systemic risks. What do you think about these options?
    a. Content warnings/hiding harmful content (ex. Twitter and more recently, Facebook)
    b. Downvoting/Upvoting content (Facebook primarily uses its algorithm to determine which content is more or less visible - downvoting/upvoting would integrate the community in this effort)

c. Using reputation systems or internal currencies for users
d. Community moderators (a greater role for users in making content moderation decisions)
e. Automation (to provide automated feedback to users who are likely breaking rules)

11. What steps does the platform take to reduce systemic risks through changing the way it conducts online advertising?
12. What is your take on existing legal regulations regarding risks to free and fair elections and human rights?
13. How could online election observation support free and fair elections and human rights?
14. Who do you think should be held accountable for any failures to prevent or mitigate risks to free and fair elections and human rights?
15. Most large social media platforms are based in the USA. How responsive are they to concerns about their influence on elections in the rest of the world?
16. Is there anything else you want to tell us?

6.4. List of interviewees

Interviewee No.	Function
1	Expert on digital and social media and legislation, former platform executive
2	Former platform employee
3	Former platform employee
4	Legal and human rights expert and scholar
5	Hate speech and legal expert
6	Researcher on human rights
7	Online disinformation researcher
8	Researcher on social dynamics of online platforms
9	Data scientist, former platform employee
10	Researcher on online political advertisement