The 2021 German Federal Election on Social Media: Analysing Electoral Risks Created by Twitter and Facebook

Johanne Kübler[1], Marie-Therese Sekwenz, Felicitas Rachinger, Anna König, Rita Gsenger, Eliška Pírková, Ben Wagner, Matthias C. Kettemann, Michael Krennerich, Carolina Ferro

Abstract

Safeguarding democratic elections is hard. Social media plays a vital role in the discourse around elections and during electoral campaigns. The following article provides an analysis 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. This article focuses on Twitter and Facebook and their roles during the German federal elections that took place on 26 September 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 and developed systematic categories for this purpose. In conclusion, we discuss how to respond to these challenges as well as avenues for future research.

Keywords: disinformation, elections, Germany, illegal content, socio-technical systems, platform governance.

1. 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 article 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) to identify significant systemic risks stemming from the operation of their platforms, and Article 27 proposes mitigation measures that these platforms should implement. 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 (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, p. Art. 25). However, what would a concrete analysis of risks of an online platform based on the proposed DSA look like in practice?

To address this question, this article proposed to conduct an external analysis of relevant electoral risks without access to internal platform data. Of course, such an analysis is very difficult. Consequently, the article was only able to conduct a much smaller version of than the risk assessment than would be legally necessary under the proposed DSA. However, we believe our work can serve as an initial demonstration of what such analysis of risks for electoral processes could look like and contribute to the debate on how to implement them in practice.

The analysis was carried out in the context of the German federal elections that took place on 26 September 2021, taking into consideration the two large online platforms mentioned previously: Twitter

[1] Johanne Kübler, WU Wien; Marie-Therese Sekwenz, TU Delft; Felicitas Rachinger and Matthias C. Kettemann, University of Innsbruck; Anna König, Rita Gsenger, University of Potsdam; Eliška Pírková, Access Now; Ben Wagner TU Delft and Inholland; Michael Krennerich, University of Erlangen; Carolina Ferro, Enabling Digital. Corresponding Author: Ben Wagner (ben@benwagner.org).

URI: https://hdl.handle.net/10125/103124
978-0-9981331-6-4

and Facebook. The article focused exclusively on 'systemic electoral risks' rather than examining other areas of systemic risks also raised by large online platforms. 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 analysis of risks 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.

2. Theoretical and methodological approach

Online platforms, such as social media site become major avenues for the distribution of information and debates on politics, especially in the context of elections. Misinformation, disinformation, and propaganda tactics are not unique to our era (Ireton et al., 2018). 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 & Makse, 2019). Another article 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 & 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).

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.
- Deletion of content.
- Ban/suspend user accounts (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 offer alternatives to these challenges (Kiesler et al., 2012). 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 behavior 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, p. 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 behavior within specific communities may be particularly effective in getting users to change their behavior (Wagner & Kubina, 2021).

2.1. Theoretical framework

This section will discuss the theoretical framework built for this article. 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 article 'systemic electoral risks'), considering the negative impacts they might have on free and fair elections.

2.2. 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, p. 2).

Similarly, in the DSA proposal, the European Commission recognizes 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 (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, p. 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 article 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." (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, p. Art. 26).

Based on the analysis of these three categories of systemic risks, this article attempts to understand the extent to which online platforms have been able to prevent these risks.

2.3. Systemic electoral risks and categories of analysis

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" (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, p. Art. 26). However, risks in the context of elections are mentioned but not detailed in the current DSA proposal. Therefore, this article intends to tailor the areas of risk proposed by the DSA for VLOPs to the context of elections in the European Union. In this article, 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 article. Thus, in this article, systemic risks are defined as primarily falling into the following three categories:

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

This article developed a codebook consisting of three distinct parts corresponding to each category mentioned above. 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 the existing categories developed by Tiedeke et al. 2020. The electoral rights subcategories were developed based work by the authors based on the state of the art. Lastly, the disinformation subcategories are based on Kapantai et al. (2021). Below, the categories created for this article to effectuate an analysis 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 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 Terms of Service (ToS), a legal document a person must agree to abide by when registering an account.

This article 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 article, 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). The right to free and fair elections is also 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. The 'electoral rights category' 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 & 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 criticize a news story or media outlet (HLG, 2018). Online platforms implement different strategies to deal with disinformation. The 'disinformation category' created for this article 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". This is because the way the two subcategories were defined was insufficiently precise in order to be able to accurately code them into the data systematically. Therefore, these subcategories were removed, leaving nine disinformation codes.

2.4. 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 article 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. For example, we were unable to access content that the platforms had previously moderated and thus was no longer publicly available. This would have been helpful to understand platforms moderation behavior and the degree to which their actions in moderating content reduced the amounts of problematic content. 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 analyze 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 (Hao, 2021; Wagner, 2020). To explore the scope of illegal and misleading content, as well as content infringing on electoral rights, this article analyzed empirical data from two VLOPs operating in Germany.

This article relied on a 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 article applied a quantitative 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;.

2.4.1. How to conduct an analysis of electoral risks

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 & Odum Institute, 2019, p. 2). Based on these coded data, the article 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 article 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.

2.4.2. Case selection

As part of the research design, this study looks at two cases of VLOPs to assess the systemic risks online platforms pose to democratic elections. Therefore, we conducted data analysis on two VLOPs. 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 analysis of electoral risks could or should be done by studying these specific platforms.

2.4.3. Data collection

This article analyzed 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 organized 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 & Burgess, 2011; Larsson & Moe, 2012; Lin et al., 2014; Shamma et al., 2009). Hashtags serve here as an indicator that a user's messages contributed to a given topic.[2] 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, p. 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.

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 randomized samples of 1101 entries each, and finally coded using the codebook previously created. The deductive codebook was based on the state of the art in the respective academic fields, and is included in the Annex. The codebook includes detailed categories in the area of dissemination of illegal content, electoral rights, and disinformation.[3] In addition, the categories we developed are not perfectly distinct and at times overlap, with multiple categories fitting to a single post. For example, one Facebook post can both be disinformation and an electoral rights violation at the same time.

Three coders were involved in the coding process. All of the legal aspects of the coding were law students who had studied law for at least 3 years. They coded the material separately using Microsoft Excel. For cases where there was a disagreement between at least two coders, an additional group of four senior researchers (with legal and social science backgrounds) came together with the coders to discuss the individual cases and mutually agree on an appropriate outcome. 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] 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.

[3] We cannot go into significant detail here about the different categories used, due to limitations in the possible length of this article. However, we have provided the full list of coding categories and supporting materials as an Annex (see Section 5).

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 our 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 analyzed. 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 article 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) 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.

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 behavior" (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, delegitimize 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 delegitimizing elections, encouraging electoral protests." (Codebook).

Moreover, 1.35% were coded as subcategory 'Procedural – Vote count' (E-8), which is defined as "Actors interested in delegitimizing 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-)legitimizing the elections or harming/promoting certain candidates or parties publish fictitious, false, or supportive election polls to (de-)mobilize 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.

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 & 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 & Derakshan, 2017), which can be styled as news articles to make them appear legitimate.

Only on the Facebook sample the article 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 et al., 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 article 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).

4. Conclusion

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 large prevalence of problematic content in our analysis suggests that online platforms are not currently doing enough to respond to the challenge of problematic content around elections.

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, p. 743), one leading election observer acknowledged, even as he spent his days "running after the tech companies." (Wagner, 2020, p. 743).

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

Bovet, A., & Makse, H. A. (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, A., & Burgess, J. (2011). #ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics & Culture*, *44*(2), 37–56.

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, COM(2020) 825 final, 2020/0361 (COD) (2020). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

Hao, K. (2021). *How Facebook got addicted to spreading misinformation*. MIT Technology Review. 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.* (European Commission. Directorate General for Communications Networks, Content and Technology., Ed.). Publications Office. https://data.europa.eu/doi/10.2759/739290

Howard, P. N., & Kollanyi, B. (2016). *Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum* (SSRN Scholarly Paper ID 2798311). Social Science Research Network. https://doi.org/10.2139/ssrn.2798311

Ireton, C., Posetti, J., & UNESCO. (2018). *Journalism, 'fake news' et disinformation: Handbook for journalism education and training*. http://unesdoc.unesco.org/images/0026/002655/265552E.pdf

Juncker, J.-C. (2018). *State of the Union 2018: European Commission proposes measures for securing free and fair European elections* [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/IP_18_5681

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, *23*(5), 1301–1326. https://doi.org/10.1177/1461444820959296

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, *18*(1).

Kiesler, S., Kraut, R., Resnick, P., & Kittur, A. (2012). Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design* (pp. 125–178). MIT Press.

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

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747. https://doi.org/10.1177/1461444811422894

Lin, Y.-R., Keegan, B., Margolin, D., & Lazer, D. (2014). Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events. *PLoS ONE*, *9*(5), e94093. https://doi.org/10.1371/journal.pone.0094093

Manheim, D. (2020). The Fragile World Hypothesis: Complexity, Fragility, and Systemic Existential Risk. *Futures*, *122*, 102570. https://doi.org/10.1016/j.futures.2020.102570

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Mihas, P. & Odum Institute. (2019). *Learn to Build a Codebook for a Generic Qualitative Study*. SAGE Publications, Ltd. https://doi.org/10.4135/9781526496058

Peterson, W. A., & Gist, N. P. (1951). Rumor and Public Opinion. *American Journal of Sociology*, *57*(2), 159–167. https://doi.org/10.1086/220916

Rhodes, S. 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, S. T. (2017). Content Moderation. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_44-1

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Rubin, V. L., Chen, Y., & Conroy, N. K. (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

Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: Understanding community annotation of uncollected sources. *Proceedings of the First SIGMM Workshop on Social Media - WSM '09*, 3. https://doi.org/10.1145/1631144.1631148

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, *34*(3), 150–160. https://doi.org/10.1177/0266382117722446

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (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, A.-S., Kettemann, M. C., Rachinger, F., Sekwenz, M.-T., & Wagner, B. (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

United Nations. (1948). *Universal Declaration of Human Rights*. United Nations; United Nations. https://www.un.org/en/about-us/universal-declaration-of-human-rights

Wagner, B. (2020). Digital Election Observation: Regulatory Challenges around Legal Online Content. *The Political Quarterly*, *91*(4), 739–744. https://doi.org/10.1111/1467-923X.12903

Wagner, B., & Kubina, M. (2021, February 18). *Ergebnisse des Forschungsprojekts zur Stärkung der Diskussionskultur—CommunityBlog—DerStandard.at › Diskurs*. https://www.derstandard.at/story/2000124046106/ergebnisse-des-forschungsprojekts-zur-staerkung-der-diskussionskultur

Wagner, B., Kübler, J., Pírková, E., Gsenger, R., & Ferro, C. (2021). *Reimagining content moderation and safeguarding fundamental rights: A study on community-led platforms* (p. 54). Enabling Digital Rights and Governance & EU Greens/EFA. https://enabling-digital.eu/wp-content/uploads/2021/07/Alternative-content_web.pdf

Wardle, C., & Derakshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

Wardle, C., Greason, G., Kerwin, J., & Dias, N. (2018). Information Disorder, Part 1: The Essential Glossary. *First Draft*. https://medium.com/1st-draft/information-disorder-part-1-the-essential-glossary-19953c544fe3

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (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