LL.M General Law (Thesis) Alessia Zornetta
AY 2020/2021 260967862

"Online misinformation: improving transparency in content moderation practices of social media companies"

Prof. Ignacio Cofone
McGill Faculty of Law

(25227 words)

August 15th, 2021

TABLE OF CONTENTS

1. Introduction

2016 was a turning point in the history of social media. The scandals concerning the US Elections and the UK Brexit Referendum, as well as the rise of radical political parties worldwide, exposed the impact that social media have on real life and how easily malicious actors can exploit their power.[1] In a time where individuals rely ever less on traditional media[2] and where social interactions are restricted, social media became, for many, the main mean by which individuals interact with one another and obtain news on a daily basis.[3] At present, one-third of the world population routinely uses social media to interact with others, engage in public discourse, gather information, and promote their ideas and beliefs.

At first, social media have been revolutionary in giving a voice to those silenced, for the very same reason why they are now being criticized: anyone can post almost anything at any time and potentially reach everyone everywhere.[4] However, over time, this absolute freedom and lack of editorial choices and filtering came with issues. The unprecedented access to the internet, the ease with which content becomes viral, and the shield of anonymity enabled

[1] See e.g. Sheera Frenkel & Mike Isaac, "Inside Facebook's Election 'War Room'", *The New York Times* (19 September 2018), online: *The New York Times* <https://www.nytimes.com/2018/09/19/technology/facebook-election-war-room.html>. Note that Facebook's approach to political misinformation will be addressed more in detail in Section 3.

[2] Felix Salmon, "Media trust hits new low" (21 January 2021), online: *Axios* <https://www.axios.com/media-trust-crisis-2bf0ec1c-00c0-4901-9069-e26b21c283a9.html> (in 2021, 54% of surveyed Americans admitted not to trust in traditional media, and 58% believed that "most news organizations are more concerned with supporting an ideology or political position than with informing the public."); Amy Watson, "Europe: trust in the written press by country 2021 Statista" (21 May 2021), online: Statista <https://www.statista.com/statistics/454403/europe-trust-in-the-written-press-by-country/> (the average trust in the written press among the EU Member States in 2021 was 51%).

[3] Elisa Shearer, "More than eight-in-ten Americans get news from digital devices", (12 January 2021), online: *Pew Research Center* <https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/> (according to the Pew Research Center, in 2020 social media were the preferred news source for 53% of surveyed Americans).

[4] It is now commonly agreed that Facebook and Twitter were fundamental in giving effect to the uprisings of the Arab Spring. Christos A. Frangonikolopoulos & Ioannis Chapsos, "Explaining the Role and the Impact of the Social Media in the Arab Spring", (2012) 8:1 GMJ: Mediterranean Edition 10, online (pdf): <https://www.academia.edu/download/30406181/Global_Media_Journal.pdf > ("social media acted as an 'accelerating agent' […] that helped protesters hold online discussions and organize and stage popular uprisings which, in turn, led to the resignation of two unpopular leaders, as well as their rapid spread across Arab countries, and transnationalization to the wider world" at 10); Philip N. Howard et al., "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" (2011), PITPI Working Paper 2011.1, DOI: <10.2139/ssrn.2595096>.

the presence of problematic behaviour on platforms, ranging from hate speech to the spread of illegal content such as terrorist incitement and child-sexual-abuse, from unregulated advertisement to political manipulation.[5]

In the past six years, national and regional governments worldwide have proposed and enacted new laws to limit the impact of this behaviour and constrain the power of social media giants like Facebook, Twitter, and YouTube. Additionally, academic networks have increased their focus on platform governance and formed research groups and independent databases to improve third-party oversight of companies' actions. Moreover, all major social media companies have progressively established new policies and procedures to counter problematic trends in their platforms.

Nevertheless, the majority of these initiatives have failed to deliver the expected outcomes due to the ambiguity and inefficiency of transparency requirements, a common denominator among them. While transparency has been long considered as an efficient means to improve companies' practices due to the increased oversight, the same cannot be argued with transparency requirements over online content moderation[6]. Indeed, content moderation practices have long been criticized for their opacity and arbitrariness. Screening algorithms have proved fallible in several circumstances and have been condemned for reflecting and encoding social biases while operating behind the shield of ineligibility. Moreover, little is known about moderators' training or the internal protocols they are required to follow. This

[5] Garth Jowett & Victoria O'Donnell, *Propaganda & persuasion* (Thousand Oaks: Sage, 2012), online (pdf): <https://hiddenhistorycenter.org/wp-content/uploads/2016/10/PropagandaPersuasion2012.pdf > ("The very "democracy" and accessibility of the World Wide Web has made it the most potent force for the spreading of misinformation yet devised" at 159).

[6] "Content moderation" is the practice by which social media companies enforce their internal policies. Every time a user registers to social media, they agree to respect community guidelines and terms of service. To guarantee that problematic content does not end up on users' feeds, all uploaded content is moderated. Every picture, video or post on social media is screened through automated machine learning algorithms that use tools such as filtering through keywords and hashing technology to determine whether it should be made visible. In some cases, however, algorithms are not able to decide whether a certain content pertains or not to a prohibited category, so human moderators get involved and have the final say. Additionally, users and fact-checkers can also flag content that algorithms do not detect to be reviewed by human moderators.

has led academics and regulators worldwide to suggest more stringent oversight and limitations on platforms.

In this work, I address how transparency practices can tackle the issue some define as the main challenge for democracy worldwide: online mis- and disinformation.[7] The spread of misleading and false information has been defined in numerous ways amongst different platforms and actors. At present, finding a commonly shared definition is still a complex task. The most common terms used to discuss the phenomenon are "disinformation," "misinformation," and "fake news."[8] Both "disinformation" and "misinformation" are at the centre of legislative proposals over platform governance, scholarly discussions, and social media policies and reforms. They address inaccurate and potentially misleading content, the main difference between the two being that the former is used when such content is spread with the intent to manipulate and influence others, whereas, in the latter, its author or sharer is not aware of its incorrectness.[9] Even though criticism has mainly focused on disinformation rather

[7] Georgios Terzis et al, *Disinformation and digital media as a challenge for democracy* (Cambridge: Intersentia, 2020); European Parliament, *Study on The impact of disinformation on democratic processes and human rights in the world*, by Carme Colomina, Héctor Sánchez Margalef & Richard Youngs (April 2021), online (pdf): *European Parliament* < https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf>; Paul Butcher, "Disinformation and democracy: The home front in the information war" (discussion paper for the European Politics And Institutions Programme, 30 January 2019) European Policy Center, online (pdf): < https://www.epc.eu/content/PDF/2019/190130_Disinformationdemocracy_PB.pdf>; Samantha Bradshaw, The Social Media Challenge for Democracy: Propaganda and Disinformation in a Platform Society, (PhD Thesis, University of Oxford, 2020) [unpublished], online: *Oxford University Research Archive* <https://ora.ox.ac.uk/objects/uuid:e75e4796-d614-454b-b2e2-df6b8659e610>; *Contra* Andreas Jungherr & Ralph Schroeder, "Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy" (2021) 7:1 Social Media + Society 205630512198892, DOI: <10.1177/2056305121988928> (who argue that the threat of online disinformation to democracy is a moral panic resulting from "ill-understood deeper structural shifts under way that give rise to unfocused fears").

[8] Although the latter quickly became popular at first, it is often used by individuals trying to discredit information they perceive as personally unfavourable even when factual. Therefore scholars, regulators, and even the industry itself have gradually abandoned its use. The controversies with the term "fake news" are further discussed in Section 4.3.2. See also European Commission, *A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation*, by Madeleine De Cock Buning (March 2018) at 5, online (pdf): *European University Institute* <http://diana-n.iue.it:8080/handle/1814/70297>.

[9] Paul Butcher, *supra* note 7 at 3. See also Barrie Sander & Nicholas Tsagourias, " The covid-19 Infodemic and Online Platforms as Intermediary Fiduciaries under International Law" (2020) 11:2 JIHLS 331 at 333, 335, online: *JIHLS* https://brill.com/view/journals/ihls/11/2/ihls.11.issue-2.xml (where the authors further distinguish between "disinformation" and "malinformation", meaning, respectively, as "o the intentional creation and/or dissemination of verifiably false or misleading information, typically by organised state or nonstate actors", and

than misinformation, some have argued that the intent with which content is shared is not relevant when assessing its impact. Recipients of misleading content have no means of being aware of the intentions of its author or sharer. Additionally, establishing intent is particularly difficult in the online environment, given the variety of languages, jargon and nuances possible.[10] Therefore, throughout this work, I use the term "misinformation" to refer to any kind of incorrect or misleading content shared online, with or without malicious intent.

I address the current criticism over companies' insufficient efforts to tackle online misinformation and how they react to regulators' demands to limit the spread of the phenomenon, given its pervasive impact on society and real-life consequences. By focusing on regulatory initiatives in different jurisdictions, I show how to improve transparency mechanisms through standardization and uniformity. Doing so will enable platforms and governments to evaluate whether current practices are efficient in limiting the spread of online misinformation and what must be done to improve such practices. Discussions worldwide are increasing over potential regulations aimed at enhancing government control and platforms' responsibility for content moderation practices through transparency. Therefore, it appears necessary to bring light to the inefficiencies of the current transparency-reliant systems and how they could be overcome.

Section 2 focuses on the role of transparency mechanisms adopted by companies either voluntarily or to comply with national and supranational legislation. I claim that the current transparency mechanisms present significant shortcomings that prevent a meaningful comparison and an evaluation of companies' actions, as well as their compliance with regulations. In Section 3, I use Facebook's handling of political misinformation as a case study to show transparency's efficient and inefficient results. I chose to focus on Facebook due to its

"the intentional creation and/or dissemination of information that is threatening, abusive, discriminatory, harassing or disruptive, which aims to cause harm to a person, organisation or state").

[10] Jason Pielemeier, "Disentangling Disinformation: What Makes Regulating Disinformation So Difficult?" (2020) 2020:4 Utah L Rev 917 at 922-923.

continuous growth in users worldwide and the unprecedented oversight it has received since 2016. I analyze the changes introduced by the company as a response to both public and government requests worldwide by focusing on a selection of political events: the 2018 Brazilian Federal Elections, the 2019 European Parliament Elections, and the 2020 U.S. Presidential Elections. Lastly, after identifying the problems concerning the current state of transparency-reliant systems, in Section 4, I propose four areas of improvement concerning disclosure of ex ante and ex post review of content; social media internal decision-making protocols; engagement with misinformation; and user awareness tools.

2. Transparency

Misinformation is a remarkably nuanced phenomenon and deciding whether a piece of false information has been spread with the intent to harm or whether there was an innocent lack of knowledge on behalf of its author is a particularly error-prone task. Nevertheless, misinformation continues to grow on social media, which has resulted in a plurality of initiatives aimed at slowing its spread.[11] Besides legislative and regulatory initiatives, all major social media platforms have committed to being more proactive and are constantly implementing new policies and protocols to counter the phenomenon.

Transparency plays a fundamental role in determining similarities and discrepancies among platforms, with the overall aim of identifying current and future trends. The disclosure of data regarding the spread of misinformation on social media can facilitate the understanding of suspicious behaviour patterns. That is, means of circulation, time of virality, country and region of origin, and users' engagement. Identifying these patterns is helpful to tackle misinformation for a variety of reasons. Firstly, it provides platforms themselves data over the

[11] There has been a specific increase in legislative action from regulators worldwide targeting misinformation, especially when it concerns political content. See e.g. Loi organique n° 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1), JO, 23 December 2018 [*French law against manipulation of information*]; Elections Modernization Act, SC 2018, c 31 [*Canada Elections Act*];

efficiencies and weaknesses of current containment systems. Secondly, it allows independent assessments of platforms' initiatives to combat misinformation, which is helpful to national and regional regulators developing anti-misinformation laws and strategies by ascertaining vulnerable areas, especially in terms of scale and scope. Thirdly, it enables users to make conscious decisions over the reliability and integrity of content viewed and shared.[12] Therefore, disclosing data concerning the results of content moderation practices also serves the general purpose of informing users and regulators of the current online behavioural trends and safety status of platforms.[13]

2.1. Transparency-mechanisms

Before analyzing the shortcomings of current transparency-reliant systems, a clarification on transparency mechanisms is needed. Transparency mechanisms have been used for decades already before the rise of online platforms. Modern transparency has been identified as a tool to maintain citizens informed and governments accountable based on the empowerment obtained through access to information.[14] With time, transparency mechanisms have also influenced private corporations, especially when their actions have a remarkable impact on individuals and the public sphere.[15] These transparency-reliant systems function on the assumption that, by making available to the public scrutiny the internal corporate decision-making, on the one hand, the customers of such companies will be able to make informed

[12] European Commission, A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation, by Madeleine De Cock Buning (March 2018) at 22, online (pdf): European University Institute <http://diana-n.iue.it:8080/handle/1814/70297>.
[13] But see Mikkel Flyverbom, "Digital Age Transparency: Mediation and the Management of Visibilities" (2016) 10:0 International Journal of Communication 13, online: <https://ijoc.org/index.php/ijoc/article/view/4490> (arguing that transparency mechanisms are tools of "visibility management" used by platforms as a means to avoid external intervention).
[14] Mikkel Flyverbom, "Sunlight in cyberspace? On transparency as a form of ordering" (2015) 18:2 European Journal of Social Theory 168 at 169, online: <https://journals.sagepub.com/doi/10.1177/1368431014555258>
[15] Don Tapscott & David Ticoll, Naked corporation: how the age of transparency will revolutionize business (Toronto: Viking Canada, 2012).

decisions on whether to keep using their services[16] and, on the other, regulators will be able to assess companies' compliance with the law and eventually promote new well-functioning regulations.[17]

Some transparency mechanisms result from proactive initiatives by private companies themselves, such as through notices to affected individuals, high-profile press releases, and transparency reports, which often provide information about the company's actions to tackle a specific issue. Others are the result of companies' compliance with laws and regulations demanding disclosure or with public filings during judicial proceedings.[18] Lastly, research made by independent auditors and documents obtained and published unofficially by the media are also considered part of the ecosystem of transparency mechanisms.[19]

Transparency mechanisms differ depending on the kind of platform transparency.[20] The primary and most popular mechanism among these is transparency reporting, i.e. when private companies engage in public disclosure. Nowadays, all major online platforms communicate information on their content moderation practices to the public either proactively or as fulfilling regulatory duties imposed by national and regional laws. In the field of internet platforms, reports include data on how much content is removed from a platform, either upon

[16] Archon Fung, "Infotopia: Unleashing the democratic power of transparency" (2013) 41:2 Politics & Society 183 at 184, online: <https://journals.sagepub.com/doi/abs/10.1177/0032329213483107>.
[17] Daphne Keller & Paddi Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily & Joshua A. Tucker, eds, *Social media and democracy: the state of the field, prospects for reform* (Cambridge, United Kingdom; New York, Ny: Cambridge University Press, 2020) 220 at 222.
[18] *Ibid*.
[19] See Jurgen de Jong & Michiel S. de Vries, "Towards unlimited transparency? Morals and facts concerning leaking to the press by public officials in the Netherlands" (2007) 27:3 Public Administration and Development 215 (who claim that, despite its negative connotation, leaking serves the public interest); Mikkel Flyverbom *supra* note 14 at 175 (where it is argued that whistleblowing and professional transparency organizations improve transparency of online services).
[20] See Robert Gorwa & Timothy Garton Ash, "Democratic Transparency in the Platform Society" in Nathaniel Persily & Joshua A. Tucker, eds, *Social media and democracy : the state of the field, prospects for reform* (Cambridge, United Kingdom ; New York, Ny: Cambridge University Press, 2020) 286 at 295 (where they subdivide platforms'' transparency mechanisms into "voluntary transparency around freedom of expression and for content takedowns; legally mandated transparency regimes; self-transparency around advertising and content moderation; and third-party tools, investigations, and audits.").

governmental request or due to a violation of the platforms' terms of services and community guidelines.[21]

For instance, Facebook started to publish quarterly reports on requests for user data and content removal in 2013. Since then, it has gradually expanded the information available on its content moderation practices.[22] Nevertheless, a particular turning point in Facebook transparency reporting dates back to 2018, when it joined the Global Network Initiative. From this moment, it started to publish a detailed explanation of the internal enforcement process of "Community Standards," which had been kept secret until then.[23] Nowadays, Facebook's transparency report is divided into three sections dealing with the enforcement of its community standards, responses to legal requests, and internet disruptions caused by external forces.[24] It is important to notice that the former is particularly valuable considering the frequency with which community standards are updated to respond to social changes and demands. At the time of writing, the latest "Community Standards Enforcement Report" (FB CSER) contains information over both Facebook and Instagram's moderation efforts in twelve categories, including illegal content[25] and harmful content[26].[27]

In addition to community standards enforcement reports, major platforms also disclose data over their content moderation practices in response to government demands. For instance, since 2010, Google has been publishing detailed biannual reports containing data over national

---

[21] *See* Facebook Inc., "Facebook Transparency Report", online: *Facebook* <https://transparency.facebook.com/>; "Twitter Transparency Center", online: *Twitter* <https://transparency.twitter.com/>; "Google Transparency Report", online: *Google* <https://transparencyreport.google.com/youtube-policy/removals>.
[22] Hayley Tsukayama, "Facebook releases first report on world governments' data requests" (27 August 2013), online: *Washington Post* <https://www.washingtonpost.com/business/technology/facebook-releases-first-report-on-world-governments-data-requests/2013/08/27/40e2d396-0f24-11e3-8cdd-bcdc09410972_story.html>.
[23] Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process - About Facebook" (24 April 2018), online: *Facebook* <https://about.fb.com/news/2018/04/comprehensive-community-standards/>.
[24]Facebook Inc., *supra* note 21.
[25] Such as child nudity and sexual exploitation of children, terrorism, and unregulated goods.
[26] Such as fake accounts, violent and graphic content, and suicide and self-injury.
[27] Facebook Inc., "Facebook Community Standards Enforcement Report" (Q1 2021), online: *Facebook* <https://transparency.facebook.com/community-standards-enforcement>.

courts and agencies worldwide requesting the removal of information on the grounds of national security, copyright violations, defamation, illegal goods and services, and privacy and security.[28] While some local laws require compliance reports to be produced by platforms, it has now become standard practice for companies to disclose data about content removal upon governmental requests voluntarily. This kind of disclosure – especially in countries governed by authoritarian regimes – often aims to provide users with a justification. In this way, the accountability for the removal is shifted to the local government rather than the company, which positions itself as a counterpart.[29] Although this activity has been criticized by some claiming that such disclosure is politically motivated,[30] it can be argued that such a transparency initiative has positive effects in providing users with information of government interference with online public discourse.[31]

Transparency reports are not always necessarily the result of platforms' voluntary initiatives, but, instead, they are also issued in response to obligations imposed by national and regional laws. For instance, the French Law on Manipulation of Information requires platforms to publish annual reports on the progress made in fighting online misinformation, including statistical information on algorithms used to rank content on users' feeds.[32] As will be discussed later, mandatory impositions are not devoid of criticism but remain a popular tool among governments trying to curb the power of online platforms.

---

[28] Google LLC, "Google Transparency Report - Government requests to remove content" (2021), online: *Google* <https://transparencyreport.google.com/government-removals/overview?hl=en>.

[29] *Cf* Eleni Kosta & Magdalena Brewczyńska, "Government Access to User Data: Towards More Meaningful Transparency Reports" in Rosa M Ballardini, Petri Kuoppamäki & Olli Pitkänen, eds, *Regulating Industrial Internet Through IPR, Data Protection and Competition Law* (Kluwer Law International, 2019) 1 at 14, online: *SSRN* < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3601661> (on companies motivation for voluntary disclosure of government user's data requests).

[30] Monika Zalnieriute, "'Transparency Washing' in the Digital Age: A Corporate Agenda of Procedural Fetishism" (2021) 8:1 Critical Analysis of Law 140 at 142.

[31] *Cf* Christopher Parsons, "The (In)effectiveness of Voluntarily Produced Transparency Reports" (2017) 58:1 Business & Society 103, online: <https://journals.sagepub.com/doi/full/10.1177/0007650317717957> (discussing the effectiveness of telecommunications companies' transparency reports)

[32] Rachael Craufurd Smith, "Fake news, French Law and democratic legitimacy: lessons for the United Kingdom?" (2019) 11:1 Journal of Media Law 52 at 62-63, DOI: <10.1080/17577632.2019.1679424>.

Lastly, another standard tool of company transparency, whose use has increased recently, is company press releases over high-profile issues. Examples of the latter often deal with disinformation during political electoral campaigns,[33] foreign and domestic terrorism attacks,[34] responses to leaked documents, and fixing of content moderation mistakes.[35] The most recent example regards the explanations issued by major social media companies which decided to suspend the accounts of the former U.S. President Donald Trump upon the alleged incitement of the Capitol Hill violence of January 6th, 2021.[36]

While the focus of this work will remain on transparency reports, it needs to be observed that other transparency mechanisms such as user notifications and labelling systems are also widely used in the fight against online disinformation. However, their actual efficacy has been the subject of controversy.[37]

2.2. When is transparency valuable in tackling online misinformation?

Even though content moderation is not new for platforms, content moderation of misinformation has been the focus of policy initiatives both within and outside companies'

[33] See e.g. Anika Geisel, "Protecting the European Parliament Elections - About Facebook" (28 January 2019), online: *Facebook* <https://about.fb.com/news/2019/01/european-parliament-elections/> (on the steps taken by Facebook during the European Parliament Elections of 2019 to combat inauthentic behaviour and misinformation).

[34] See e.g. Guy Rosen, "Protecting Facebook Live From Abuse and Investing in Manipulated Media Research - About Facebook" (15 May 2019), online: *Facebook* <https://about.fb.com/news/2019/05/protecting-live-from-abuse/> (on Facebook's response to the Christchurch Mosque Shootings of March 15th, 2019, which were livestreamed on its platform).

[35]See Abby Ohlheiser, "Facebook backs down, will no longer censor the iconic 'Napalm Girl' war photo"(9 September 2016), online: *Washington Post* <https://www.washingtonpost.com/news/the-intersect/wp/2016/09/09/abusing-your-power-mark-zuckerberg-slammed-after-facebook-censors-vietnam-war-photo/> (on Facebook's response to its wrongful removal of the Pulitzer-prize winning image "Napalm Girl" by Nick Ut); Kimberly Hall, "Public Penitence: Facebook and the Performance of Apology" (2020) 6:2 Social Media + Society 205630512090794, online: <https://journals.sagepub.com/doi/full/10.1177/2056305120907945>.

[36] Guy Rosen & Monika Bickert, "Our Response to the Violence in Washington - About Facebook" (7 January 2021), online: *Facebook* <https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/>; Twitter Inc., "Permanent suspension of @realDonaldTrump", (8 January 2021), online: Twitter <https://blog.twitter.com/en_us/topics/company/2020/suspension.html>.

[37] See e.g. Megan A Brown et al, "Twitter put warning labels on hundreds of thousands of tweets. Our research examined which worked best.", *the Washington Post* (9 December 2020), online: <https://www.washingtonpost.com/politics/2020/12/09/twitter-put-warning-labels-hundreds-thousands-tweets-our-research-examined-which-worked-best/>.

headquarters. In the past decade, the spread of misinformation online has been facilitated by the new "marketplace of ideas" created by social media.[38] Online platforms give unprecedented visibility to anyone able to create a post worthy of attracting the attention of one's network. Moreover, social media business models (which depend on users' engagement) rely on algorithms that reinforce the visibility of misinformation posts, which are often formulated to become viral.[39]

The impact of online misinformation on real-life events has moved the spotlight to platforms' role in shaping public opinion, which inevitably results in increased demands for accountability.[40] However, accountability and public scrutiny inherently depend on the ability of the public to obtain information about platforms' operations, especially in terms of policy- and decision-making.[41]

Governments worldwide have proposed more stringent measures to limit the power of and incentivize platforms' accountability over content moderation practices. As of 2020, examples of successful and failed attempts could be found in fifty-three countries worldwide.

---

[38] Tim Hwang, "Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment" (2017) SSRN Electronic Journal at 39, online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442> [unpublished].

[39] Laura Elderson et al, "Far-right news sources on Facebook more engaging" (3 March 2021), online (Medium): *Cybersecurity for Democracy* <https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90> (who analysed user engagement on Facebook during the 2020 US elections and found that "[f]ar-right sources designated as spreaders of misinformation had an average of 426 interactions per thousand followers per week, while non-misinformation sources had an average of 259 weekly interactions per thousand followers"); Flavia Durach, Alina Bârgăoanu & Cătălina Nastasiu, "Tackling Disinformation: EU Regulation of the Digital Space" (2020) 20:1 Romanian Journal of European Affairs 5 at 7, online: <https://www.ceeol.com/search/article-detail?id=859431> (where the author argues that disinformation stimulates engagement); *and* James Grimmelmann, "The Platform is the Message" (2018) 2 Geo L Tech Rev 217 at 227, online: Georgetown Law and Technology Review <https://georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Grimmelmann-pp-217-33.pdf>.

[40] The platforms' essential role in shaping public discourse has been repeatedly stressed by scholars. See e.g. Jack M. Balkin, "Free Speech is a Triangle" (2018) 118:7 Colum L Rev 2011, online: *Columbia Law Review* <https://columbialawreview.org/content/free-speech-is-a-triangle/> (who defines digital companies such as Facebook and Google as "custodians of the public sphere and of democratic self-government" at 2041); *and* Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2017) 131 Harv L Rev 1598, online: <https://ssrn.com/abstract=2937985> (who argues that "[d]igital speech has created a global democratic culture, and the [platforms] are the architects of the governance structure that runs it" at 1664).

[41] Michael Karanicolas, "A FOIA for Facebook: Meaningful Transparency for Online Platforms" (Paper delivered at the FESC of the Floyd Abrams Institute for Freedom of Expression, Yale Law School, 30 April 2021) [unpublished]. *See also* Archon Fung, Mary Graham & David Weil, *Full disclosure: the perils and promise of transparency* (New York: Cambridge Univ. Press, 2010).

While not all proposed interventions can be said to have as their primary objective that of preventing the spread of online misinformation, some jurisdictions were successful in introducing reporting duties and forced online platforms to strengthen and improve their policies dealing with the phenomenon. Nevertheless, it is essential to note that these measures differ in scope and application, ranging from soft-law instruments covering misinformation broadly[42] to mandatory laws targeting only particular, complex, and vulnerable circumstances such as election campaigns[43] or, more recently, the Covid-19 pandemic.

Social media companies have also recognized the power of transparency in tackling critical issues inside their platforms. It is undeniable that most transparency initiatives enacted in the past three years had as initial aim that of restoring platforms' reputation after the criticism received following the Cambridge Analytica scandal.[44] However, disclosure of the company's practices and decision-making has also provided platforms with an opportunity to demonstrate reactiveness and readiness to address complex issues and improve the quality of users' experience.

Transparency mechanisms force platforms to enhance their actions. Once companies disclose information about their content moderation practices, they put themselves under public scrutiny and are, therefore, pressured to improve their achievements over time. Indeed, it could

[42] An example of a soft law instrument is the EU Code of Practice on Disinformation, which has been voluntarily signed by all major online companies operating in the European Economic Area. Its signatories have agreed to report monthly on their actions to counter the spread of misinformation, to increase media literacy tools available to users, and to collaborate with independent third parties both in ensuring the accuracy of information shared on their platforms and in evaluating the efficiency of companies' actions. Nevertheless, the voluntary nature of the Code and its limited application render its obligations less stringent. See more in Section 3.2.

[43] For instance, the Canadian Elections Modernization Act demands that platforms make available an online registry of partisan and election advertisement messages for two years after the elections. Furthermore, it prohibits the dissemination of false information about candidates aimed at influencing the elections, and the purchase of regulated ads by foreigners and foreign entities during the election period. See Canada Elections Act, *supra* note 11 at ss 91-92, 208.1, 349 (1)(b).

[44] Nicholas Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far (Published 2018)", *The New York Times* (4 April 2021), online: <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>; Daniel Susser, Beate Roessler & Helen F. Nissenbaum, "Online Manipulation: Hidden Influences in a Digital World" (2018) 4:1 Georgetown Law Technology Review, DOI: <10.2139/ssrn.3306006>;

be argued that companies that have not disclosed data over the actions taken to counter problematic behaviour on their platform did not achieve long-term trust and inevitably failed. The most recent example of such failure is Parler – a micro-blog platform similar to Twitter – which was at the centre of discussions in the aftermath of the storming of Capitol Hill.[45] By not adopting policies and protocols able to counter the spread of misinformation, individuals who found obstacles in major platforms like Facebook and Twitter moved to Parler and used it to spread misinformation and coordinate the violence.[46] As a result, Parler was removed from Amazon's cloud computing servers and banned by both Apple and Google in their app stores. This resulted in the loss of its business partnerships and thus any sort of income, which inevitably led the at-the-time Parler's CEO, John Matze, to announce the shutdown of the platform.[47]

Furthermore, transparency over content moderation practices enables social media companies such as Facebook and Twitter to demonstrate their willingness to cooperate with regulators worldwide, which is often a means to avoid more stringent regulation.[48] At the same time, it can also be beneficial to users and companies when dealing with excessive government intervention. Indeed, while many in democratic countries see government intervention as a viable alternative to limit the power of platforms in shaping public opinion,[49] it can have dangerous effects in non-democratic countries. This is because platforms often represent the

[45] David Shepardson, "U.S. panel asks FBI to review role of Parler in Jan. 6 Capitol attack", (21 January 2021), *Reuters* online: <https://www.reuters.com/article/us-usa-trump-parler-idUSKBN29Q2FS>.
[46] Mike Isaac & Kellen Browning, "Fact-Checked on Facebook and Twitter, Conservatives Switch Their Apps", *The New York Times* (11 November 2021), online: <https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html>.
[47] Elizabeth Culliford & Jeffrey Dastin, "Parler CEO says social media app, favored by Trump supporters, may not return", *Reuters* (13 January 2021), online: <https://www.reuters.com/technology/exclusive-parler-ceo-says-social-media-app-favored-by-trump-supporters-may-not-2021-01-13/>.
[48] Mikkel Flyverbom, *supra* note 14 at 178-179; Monika Zalnieriute, "'Transparency Washing' in the Digital Age: A Corporate Agenda of Procedural Fetishism" (2021) 8:1 Critical Analysis of Law 140 at 142.
[49] It needs to be noted that laws against misinformation have also been criticized in democratic countries on freedom of speech and the press grounds. See e.g. Rim-Sarah Alouane, "Macron's Fake News Solution Is a Problem" (29 May 2018), online: *Foreign Policy* <https://foreignpolicy.com/2018/05/29/macrons-fake-news-solution-is-a-problem/> (discussing the French Law against manipulation of information during the electoral period).

only means for government opponents to expose their oppression. In particular, this is relevant when non-democratic regimes justify regulation and operations limiting online speech behind the excuse of the fight against misinformation but instead aim at targeting and suppress their opponents.[50] Transparency mechanisms can thus also shed light on oppressive government interference, both through disclosure of government-led misinformation campaigns and government requests for take-down of content and user data. Through such disclosure, companies can place themselves against the non-democratic governments and improve their overall reputation by gaining support from the international community.[51]

Transparency can only be valuable to tackle misinformation on social media if it is *meaningful*. When the available information provides details on how decisions are taken and how platform policies are enforced beyond mere statistical data, regulators, academics, civil society, and the public can be said to have the tools necessary to evaluate platforms' actions and react accordingly.[52] Additionally, when meaningful, the information obtained through transparency mechanisms can propel policy creation and reform advancements, both at the industry and governmental levels. This is because disclosure empowers receivers of such information to hold the disclosers accountable by changing their behaviour, forcing the latter to change their own decision-making.[53]

2.3. Shortcomings of current transparency mechanisms

[50] See e.g. Akta Antiberita Tidak Benar 2018, (Federal Government Gazette, Adopted on 11 April 2018), online (pdf): https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/106305/130354/F-927153343/MYS106305%20Mys.pdf [Malaysian Anti-Fake News Act 2018]. The law criminalized the sharing of misinformation with imprisonment of up to six years and a monetary fine of up to 500,000 ringgit. It has been widely criticized for its vagueness, which had been exploited by the local government to control speech online.
[51] Mikkel Flyverbom, *supra* note 14 at 178.
[52] James Grimmelmann, "The Virtues of Moderation" (2015) 17 Yale JL & Tech 42 at 66.
[53] Archon Fung, Mary Graham & David Weil, *Full disclosure: the perils and promise of transparency* (New York: Cambridge Univ. Press, 2007) at 59-63. [*Fung et al*]

To tackle misinformation transparency needs to be meaningful.[54] At present, however, it cannot be said that the transparency mechanisms put in place by platforms provide enough meaningful information to be considered adequate and efficient in tackling online misinformation – a trend that is increasing rather than decreasing, despite companies' actions.[55] This section addresses how current transparency mechanisms fail to provide information that enables users and regulators to evaluate and hold platforms accountable.

The shortcomings mentioned below make it clear that transparency does not achieve its aims without establishing commonly shared criteria and procedures and without improving independent oversight of companies' actions. The discrepancies among platforms' transparency practices demonstrate that without specifying the terminology, metrics, categories, and legibility of data to be used, transparency practices fail to be valuable in solving the ongoing crisis of online misinformation.[56]

2.3.1. Transparency reports

Transparency reports have been criticized for being a "market-friendly" way for platforms to address critical issues and avoid regulatory oversight.[57] One of the main criticisms regards the ineligibility of the data provided by platforms that tend to disclose information

[54] Mike Ananny & Kate Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" (2016) 20:3 New Media & Soc'y 973 at 978, online: *Sage Journals* <https://journals.sagepub.com/doi/full/10.1177/1461444816676645>.
[55] Sara Fischer, "'Unreliable' news sources got more traction in 2020" (22 December 2020), online: *Axios* <https://www.axios.com/unreliable-news-sources-social-media-engagement-297bf046-c1b0-4e69-9875-05443b1dca73.html> ("In 2020, nearly one-fifth (17%) of engagement among the top 100 news sources on social media came from sources that NewsGuard deems generally unreliable, compared to about 8% in 2019")
[56] *Cf* Katharine Dommett, "Regulating Digital Campaigning: The Need for Precision in Calls for Transparency" (2020) 12:4 Policy & Internet 432 at 433, 442 (where the author argues for the need to increase specificity in transparency requirements over digital campaigning). See also Ben Wagner, *Global Free Expression - Governing the Boundaries of Internet Content* (Springer International Publishing, 2016) at 11; EU, *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*, (Commission SWD 180) (10 September 2020) at 10, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212>. [*Assessment CoP*]
[57] Nicolas P. Suzor et al, "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation" (2019) 13 Int'l J of Comm 1526 at 1529, online: <https://ijoc.org/index.php/ijoc/article/view/9736>.

about their content moderation practices through aggregated data, providing a comprehensive overview of their approach in specific areas of concern.

For example, Facebook publishes a "Community Standards Enforcement Report" every three months.[58] For each chapter of Facebook's Community Standards, data is disclosed over five criteria: i) prevalence, ii) content actioned, iii) proactive removal, iv) appealed content, and v) restored content.[59] In addition to it, since 2018, Facebook publishes a monthly report targeting "Coordinated Inauthentic Behavior" (FB CIB). This is the term the company uses to describe "coordinated efforts to manipulate public debate for a strategic goal where fake accounts are central to the operation."[60] The report is the most detailed disclosure of content moderation practices of online misinformation so far. It contains information concerning Facebook and Instagram's efforts to address the issue and is divided by source (i.e., non-government and government actors) and country.

In comparison, Twitter publishes a bi-annual report over "Twitter's Rules Enforcement," where it provides comprehensive data over three categories: i) accounts actioned, ii) accounts suspended, and iii) content removed. Instead of separate sections for each rule, Twitter's report contains one unique table comparing content moderation operations of different rules and infographics comparing each category to the previous reports to highlight increases and decreases in trends.[61] Moreover, Twitter also publishes a bi-annual "Platform Manipulation Report" with data concerning its action to counter attempts to "mislead others and/or disrupt their experience by engaging in bulk, aggressive, or deceptive activity" not

---

[58] Facebook Inc., *Facebook Transparency Report*s (2021), online: *Facebook* <https://transparency.facebook.com/> [*FB CSER*]. The report is divided in sections concerning chapters of Facebook's "Community Standards". In each section, the company provides a brief summary of the rules being enforced and of the trends perceived in comparison to previous quarters. The remaining of the report is structured through self-posed questions that the company answers with graphic data.

[59] *Ibid.*

[60] Facebook Inc., *April 2021 Coordinated Inauthentic Behavior Report* (May 2021) at 2, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2021/05/April-2021-CIB-Report.pdf>

[61] Twitter Inc., "Rules Enforcement - Report" (11 January 2021), online: *Twitter* <https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun>.

originating from government actors.[62] For actions attributable to the latter instead, Twitter provides brief updates through blog posts.

In its quarterly reports, YouTube also makes available data over the removal of channels, videos, and comments made on its platform. While for reductions of channels, clarifications are made only with regards to the "reason for removal," data over the removal of comments is also sub-divided by "source of first detection." For what concerns the removal of videos, YouTube additionally provides clarifications in terms of "views before removal" and "country/region of upload."[63]

It can be observed that the reports differ to a great extent. Firstly, platforms are inconsistent in the terminology used to address the phenomenon of misinformation. On Facebook, it is covered under different titles of its Community Standards: "inauthentic behaviour", "false news", and "manipulated media".[64] Twitter Rules, instead, deal with misinformation under three headings: "platform manipulation", "synthetic and manipulated media," and "civic integrity."[65] YouTube Community Guidelines prohibit misinformation under the title deceptive practices, which are further specified as "incitement to interfere with democratic processes" and "manipulated media."[66]

It has been argued that the use of different metrics by platforms makes it impossible to effectively compare results among them in empirically sound ways,[67] thus undermining the

[62] Twitter Inc., Platform Manipulation Report (11 January 2021), online: *Twitter* <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun>.
[63] Google LLC, "YouTube Community Guidelines Enforcement" (2021), online: Google < https://transparencyreport.google.com/youtube-policy/removals?hl=en>.
[64] Facebook Inc., "Facebook Community Standards" (2021), online: *Facebook* <https://www.facebook.com/communitystandards/>. [*FB CS*]
[65] Twitter Inc., "Twitter Rules and Policies", (2021), online: *Twitter* <https://help.twitter.com/en/rules-and-policies>. [*Twitter Rules*]
[66] Google LLC., *YouTube Community Guidelines & Policies* (2021), online: *YouTube* <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/>.
[67] Amélie Heldt, "Reading between the lines and the numbers: an analysis of the first NetzDG reports" (2019) 8:2 Internet Policy Review 1 at 11.

overall efficacy of transparency mechanisms introduced voluntarily by platforms.[68] Indeed, while platforms have constantly updated their policies to clarify the meaning of these terms further and provide practical examples of breaches, transparency reports do not reflect the same level of specificity. For instance, Twitter, Facebook, and YouTube all limit disclosure of data by choosing broad and generic headings and by dispersing information throughout different reports in the following ways. Twitter Rules Enforcement Report only provides data over "civic integrity," not specifying whether data over "platform manipulation" and "synthetic and manipulated media" are also comprised under such heading.[69] FB CSER heading "integrity and authenticity" only provides data about spam and fake accounts. To obtain information about inauthentic behaviour, auditors are referred to the monthly FB CIB. However, it only provides information about coordinated misinformation campaigns by foreign entities. YouTube's bi-monthly reports cover misinformation under the same heading as spam and scams, which leaves readers to wonder that proportion of removed content concerns misinformation specifically.[70]

These examples show that by adopting different terminologies within the same platforms and by using such language in incoherent ways throughout transparency mechanisms, companies create additional obstacles to users, researchers, and regulators trying to understand the efficiency of platforms' actions.

Secondly, these terminology discrepancies weaken the degree of specificity of the information disclosed, impeding the identification of gaps and trends within the broader umbrella of the misinformation phenomenon. By using the unique headings of "Spam" and

[68] *Cf* Ariadna Matamoros-Fernández, "Encryption poses distinct new problems: the case of WhatsApp", in Tarleton Gillespie et al, "Expanding the debate about content moderation: scholarly research agenda in the coming policy debates" (2020) 9:4 Internet Policy Review at 7-8 (where the author discusses how Facebook claims of progress in content moderation of WhatsApp "cannot be validated by external research, since only WhatsApp has access to behavioural patterns on the app").
[69] Twitter Inc., *supra* note 61.
[70] Google LLC., *Google Transparency Report* (2020), online: *Google* <https://transparencyreport.google.com/youtube-policy/removals>.

"Fake Accounts," Facebook's transparency reports do not further disclose how much content violated the manipulated media standard and how much content violated the false news standard instead. Considering that the consequences for breaching the two rules differ – content in breach of the former is removed, whereas false news are only de-prioritized – the report fails to provide meaningful information over the companies' content moderation practices. Moreover, no data is provided concerning the engagement that removed and de-prioritized content has received, thus hindering the evaluation of the impact caused by misinformation.

Thirdly, although all major platforms provide information about the content that has been removed through their content moderation operations, data about cases of over-removal or non-detection tends to be hidden or unavailable.[71] While FB CSER provides data about the amount of content appealed and restored, it needs to be taken into consideration that users still perceive the practice of appealing a decision as overly burdensome and complex, often lacking proper review.[72]

Fourthly, neither of the reports analyzed provide information about companies' efforts concerning the full extension of the phenomenon. As will be explained later, transparency reports tend to provide a positive image of companies' actions by not disclosing neither the amount of problematic content that their algorithms and moderators miss nor the engagement that deleted content received prior to removal.[73]

[71] Ben Wagner et al, "Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act" (2020) FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 261 at 266, online: <https://dl.acm.org/doi/abs/10.1145/3351095.3372856>.
[72] Sarah Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms" (2018) 20:11 New Media & Society 4366 at 4376, online: <https://journals.sagepub.com/doi/full/10.1177/1461444818773059>.
[73] See Section 3.1, *above*. See also Mikkel Flyverbom, *supra* note 14 (where the author argues that "decisions about what to make transparent and what to keep opaque can be understood as very strategic attempts at positioning organizations vis-à-vis others" at 177).

Some have also noted that disclosing aggregated data over the removal of content alone is not sufficient in ensuring the efficacy of transparency mechanisms.[74] Indeed, a commonality among all the transparency reports mentioned above is the opacity concerning the enforcement process of community standards.[75] For instance, none of the platforms at present provide accurate and verifiable information about who evaluated content's compliance with community standards[76] or what internal review procedure was followed.[77] Not disclosing this information further hinders the identification of possible shortcomings in platforms' operations. Understanding the accuracy of moderation algorithms, as well as the impact of human review of content, could provide valuable insights into the improvements needed within platforms, both in terms of algorithm efficiency and of human moderators capacity and training, also empowering regulators and users to hold platforms accountable in case of insufficient actions.

Another recurring problem of transparency reports is that the data provided by platforms have no means of being verified neither by governments nor by independent third parties.[78] By not providing information over the specificity of the content removed or restored, it is particularly difficult to evaluate the platform's accuracy in enforcement procedures since the only information available is the platform's own assessment.[79] The opacity generated around the integrity of disclosed data is damaging to the overall aim of tackling online misinformation. Firstly, it hinders the identification of lacunas in companies' policies by external observers. Secondly, it impedes an authentic evaluation by governments of the accuracy of companies' disclosures, which could result in more aggressive reactions imposing

[74] Ben Wagner et al, "Auditing Big Tech: Tackling Disinformation and the EU Digital Services Act" (2021) 1 at 8, online (pdf): *Enabling Digital Rights and Governance* <https://enabling-digital.eu/wp-content/uploads/2021/02/Auditing_big_tech_Final.pdf>; Nicolas P. Suzor, *supra* note 57 at 1525.
[75] Robert Gorwa & Timothy Ash, *supra* note 20 at 302.
[76] I.e., whether it was a human moderator or an algorithm.
[77] I.e., whether it was a first-degree review or whether the issue had been escalated. See Daphne Keller & Paddi Leersen, *supra* note 17 at 230.
[78] Ben Wagner et al, *supra* note 74 at 11; Madeleine de Cock Buning, *supra* note 8 at 14.
[79] Daphne Keller & Paddi Leersen, *supra* note 17 at 221, 228.

direct access to platforms' data, especially when misinformation campaigns target electoral processes, national security issues, and health crises.[80]

2.3.2. Compliance Reports

Compliance reports are the result of companies' compliance with mandatory reporting duties imposed by national and supranational laws. Nevertheless, most of such laws have so far been criticized for their inefficiency. For instance, the French Law Against Manipulation of Information[81] has been opposed by commentators since its draft proposal for its too broad definitions.[82] Singapore's law on the "Protection From Online Falsehoods and Manipulation" has been condemned for its far-reaching scope due to its ambiguous terminology.[83] Indeed, in the same way as voluntary transparency reports, both companies and governments can exploit this lack of specificity. In compliance reports, this enables companies to use such ambiguity in their favour by limiting the quantity and quality of information disclosed to what suits their reputational and legal needs.[84]

One of the most controversial examples of compliance reports is the disclosure requirement under the German Network Enforcement Act of 2017 (NetzDG), which demands companies to provide detailed information over their content moderation practices.[85] Accordingly, platforms with more than two million users in Germany – and meeting the

---

[80] Ben Wagner et al, *supra* note 17 at 12 (explaining why direct access in not a viable alternative by using Chinese government's direct access to data of platforms such as Sina Weibo, TikTok and WeChat).
[81] Loi organique n° 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1), JO, 23 December 2018 [French law against manipulation of information].
[82] Angelique Chrisafis, "French MPs criticise 'hasty and ineffective' fake news law", *the Guardian* (8 June 2018), online: <https://www.theguardian.com/world/2018/jun/07/france-macron-fake-news-law-criticised-parliament>.
[83] *Protection from Online Falsehoods and Manipulation Bill*, (Government Gazette, Notification No. B 10, 01 April 2019). See also Jon Russell, "Singapore passes controversial 'fake news' law which critics fear will stifle free speech" (9 May 2019), online: *TechCrunch* <https://techcrunch.com/2019/05/09/singapore-fake-news-law/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cucG95bnRlci5vcmcv&guce_referrer_sig=AQAAAFKgB98Yk4gi8GIgvBKxwY6gkpTtbuIRwI-TDoR_L_ozZAgt7jGiH6TjgJ3Ket9LFKOnQPRuRgir0JFLu3uMhL5Ui1YBDRsveDdGIcPK1UvCe1WFQ3l7E4ernf9QEmlDyUlVgLaO8BTT-CkW2XvXjQs7_oM3159KvldMRkAwwM4M>.
[84] Katharine Dommett, *supra* note 56 at 446-447.
[85] *Netzdurchsetzunggesetz* (Network Enforcement Act), Federal Law Gazette I, 3352 (2017) (Germany). [*NetzDG*]

minimum threshold of 100 notifications per month – have to publish a bi-annual transparency report in German containing, inter alia, "general observations outlining the efforts undertaken by the provider of the social network to eliminate criminally punishable activity on the platform" and the "number of complaints in the reporting period that resulted in the deletion or blocking of the content at issue."[86]

Even though the German initiative started with the aim of countering the spread of illegal content, the report is relevant to this work due to its implementation failure, which has attracted worldwide attention. Indeed, the ambiguity with which the German regulator established transparency requirements has allowed companies to interpret the obligations in their favour.[87] For instance, Facebook has exploited this ambiguity by establishing a two-step system whereby content is screened first in light of Community Standards and, only afterwards, of the NetzDG. Considering that most of the content limitations imposed by the NetzDG are also present in Facebook's Community Standards, the company was able to limit the scope of mandatory transparency under German law by only reporting content that was only removed during the second screening.[88]

Additionally, criticism over compliance reports has also focused on the fact that data provided by platforms have no means of being verified by regulators. Even though government access to data is particularly controversial, especially in circumstances of non-democratic regimes,[89] it cannot be claimed that unverified data provided by platforms should be blindly relied on when assessing compliance with legal obligations.

---

[86] §2 *NetzDG*, *supra* note 85.
[87] Ben Wagner & Carolina Ferro, *Governance of Digitalization in Europe A contribution to the Exploration Shaping Digital Policy - Towards a Fair Digital Society?,* 1st ed, Barbara Serfozo, ed (Gütersloh, Germany: Bertelsmann Stiftung, 2020) at 19.
[88] Ben Wagner, *supra* note 56 at 266. It needs to be noted that Facebook was fined by the German Federal Office of Justice. Facebook has appealed the ruling.
[89] See e.g. Min Jiang & King-Wa Fu, "Chinese Social Media and Big Data: Big Data, Big Brother, Big Profit?" (2018) 10:4 Policy & Internet 372 (on Chinese government's direct access to TikTok, WeChat and Sina Weibo).

Some regulators have tried to overcome this impasse but were inevitably unsuccessful. For instance, the European Code of Practice on Disinformation requires that platforms' data is verified by an independent third-party organization chosen by the platform itself.[90] Nevertheless, the recent assessment of the Code has found that platforms "were unable to engage an appropriate [third-party organization] and thus give effect to even this modest degree of outside review."[91] Although upon complaints from regulators, scholars, and civil society, platforms have introduced protocols to allow access to data to selected researchers for verification purposes, these have been criticized for being discretionary and excessively burdensome. This shows that, even when voluntarily committing to allow access to data, platforms companies tend to be resistant in taking proactive measures to establish such collaboration.[92] The primary purpose of independent verification of companies' disclosures is to provide an objective evaluation devoid of conflict of interests. However, when companies prevent it, such assessment and scrutiny of companies' efforts are less accurate, not only regarding compliance with mandatory duties but also within the overall scope of the misinformation phenomenon.[93] Therefore, the value of this kind of transparency mechanism is diminished when requirements are ambiguous enough to give space to interpretations that limit the scope, quantity, and quality of the information provided.[94]

2.3.3. User Notifications

[90] European Union, *Code of Practice on Disinformation*, 2018, art 20, online (pdf): <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454>. See also Section 3.2. The organization is entrusted with the evaluation of the progress made towards the commitments undertaken by platforms when signing the Code.
[91] *Assessment CoP*, *supra* note 56 at 18.
[92] Ben Wagner & Carolina Ferro, *supra* note 87 at 28. A highly controversial example is the Facebook-Harvard initiative "Social Science One" aimed at allowing researchers access to its data, where researchers have remarked that such access was actually dependent on the approval of "data grants" which were extremely limited in number. See Anja Bechmann, "Tackling Disinformation and Infodemics Demands Media Policy Changes" (2020) 8:6 Digital Journalism 855 at 857.
[93] Timothy Ash, Robert Gorwa & Danaë Metaxa, *GLASNOST! Nine ways Facebook can make itself a better forum for free speech and democracy* (Reuters Institute, 2019) at 18, online (pdf): *Reuters Institute* <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Garton_Ash_et_al_Facebook_report_FINAL_0.pdf>.
[94] Robert Gorwa and Timothy Ash, *supra* note 20 at 291.

Another transparency mechanism used by platforms is user notifications, which are often provided to the creator of content that has been removed. It is common practice for affected users to receive a communication of the removal of uploaded content due to a breach of community guidelines. For instance, Facebook will warn users that their account has been disabled when trying to log in.[95] Nevertheless, a common problem concerning user notification is the lack of a meaningful explanation over their specific case. In most platforms, users do not receive a detailed explanation of what happened to their content and what specific part of it violated community guidelines. Especially considering the complexity of identifying a post as misinformation, not offering users a personalized explanation of why their content has been taken down or why their profile has been suspended/banned could result in frustration and distrust by users. [96]

Moreover, users receive no notification of having breached policies such as FB "false news" and Twitter "manipulated media" when the remedies do not involve the deletion of such content or suspension of the account, but rather limit its distribution and reach or attach a specific label to it. Nevertheless, data regarding such actions is not disclosed in transparency reports.

This is particularly troublesome for three main reasons. Firstly, it does not allow users to contest the decision, which is problematic considering that the difficulties in identifying misinformation could lead to moderation mistakes. Secondly, not disclosing data concerning de-prioritized content in companies' transparency reports prevents an overall assessment of the spread of misinformation on their platforms. Thirdly, companies remain opaque in disclaiming

[95] Facebook Inc., "Disabled Accounts" (2018), online: *Facebook Help Center* <https://www.facebook.com/help/185747581553788>.
[96] Sarah Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms" (2018) 20:11 New Media & Society 4366 at 4376, online: <https://journals.sagepub.com/doi/full/10.1177/1461444818773059>.

the decision-making that led to applying a specific label to content, which leads to such a decision being seen as unilateral and politically biased.[97]

Similar problems regarding user notifications also concern appeal processes. All platforms admit that their practices are imperfect and could result in false positives, which is why all provide users with the possibility of appealing a removal or suspension decision. Nevertheless, studies show that appeal processes and outcomes lack as much clarity as the initial decisions. Users have complained that rather than effective reviews, appeals were merely a repetition of the initial decision without any additional scrutiny.[98]

It follows that this transparency mechanism is not sufficiently detailed to enable users to understand whether the assessment of the platform was correct or whether a mistake has been made. The opacity around reasons provided to users for content removal and de-prioritization results in a distrust over the platform's content moderation process. It also leads users to attribute the reason for removal to algorithmic and human bias.[99] Moreover, it deprives users of the opportunity to learn from their mistakes and understand what is permitted under community guidelines.[100]

Lastly, other users of the platform learn about the removal of a post through explanatory labels which appear when searching that specific content.[101] However, users who had previously engaged with removed content are generally not informed of the decision, which, in the case of disinformation, prevents them from "breaking the chain" of such content. Indeed, it has been proved that when aware of the falseness of information, users tend to refrain from

---

[97] Pranav Dixit, "Crucially, however, Twitter refused to respond to questions I asked …" (21 May 2021 06:16), online: *Twitter* <https://mobile.twitter.com/PranavDixit/status/1395730277981294598>.
[98] Sarah Myers West, *supra* note 27 at 4379. This is also reflected in transparency reports, see e.g. *FB CSER*, *supra* note 58 (according to which between April and July 2019, only 5.2 M content was restored upon users appeal against the 19.6 M requests received).
[99] Sarah Myers West, *supra* note 27 at 4373.
[100] Sarah Myers West, *supra* note 27 at 4379.
[101] *See e.g.* Twitter Inc., "Twitter account notices and what they mean - suspensions and more", (20 April 2021), online: *Twitter Help Center* <https://help.twitter.com/en/rules-and-policies/notices-on-twitter> (where it is explained that a label stating that "this tweet violated Twitter rules" will be added to tweets found in breach of community standards).

engaging with it and sharing it amongst their friends and family. Nevertheless, in this regard, advancements have been made only recently, during the COVID-19 pandemic, when regulators and health authorities demanded more action from the platforms.[102] This lack of notification to impacted users hinders the aim of countering the effects of misinformation on society since it obstacles users' overall awareness of the phenomenon.

3. Case-Study: Facebook's handling of political disinformation

Transparency mechanisms have evolved in response to social and political demands. In 2016, two political events marked a significant change in the perception of social platforms: the US Presidential Elections and the Brexit Referendum. In the aftermath of the election of Donald Trump for the US Presidency and the win of the Leave in the UK, studies demonstrated that social media had played a significant role in influencing public opinion towards these choices.[103]

Commentators agree that these revelations led to a significant shift in platforms' approach to content moderation. Analyzing the changes in Facebook's action to counter online misinformation implemented since then is helpful to understand the impact and shortcomings of transparency mechanisms on platforms' policies and regulators' demands. The following analysis of the policy changes and new features introduced by the company throughout three major political events – the 2018 Brazilian Federal Elections, the 2019 European Parliament

[102] Since April 2020, Facebook has been notifying users that have interacted with misleading content and referring them to authoritative sources of information.[102] However, this practice is – at present – only applied by Facebook and remains limited to misinformation related to COVID-19. See Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19" (16 April 2020), online: *Facebook* <https://about.fb.com/news/2020/04/covid-19-misinfo-update/>.

[103] Craig Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook", (16 November 2016), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> [showing that "top fake election news stories generated more total engagement on Facebook than top election stories from 19 major news outlets combined"]; Freedom House, *Freedom on the Net 2017: Manipulating Social Media to Undermine Democracy*, by Sanja Kelly et al. (February 2017), online: *Freedom House* < https://freedomhouse.org/report/freedom-net/2017/manipulating-social-media-undermine-democracy>.

Elections, and the 2020 US Presidential Elections – demonstrates how improving the quality of data disclosed is necessary to identify gaps in policies and regulations and develop efficient solutions.

3.1. 2018 Brazilian Presidential Elections

The Brazilian Presidential Elections of 2018 were Facebook's first opportunity to promote its increased efforts in limiting the spread of misinformation on its platforms and ensuring the authenticity of the democratic electoral process. This was not only because it was the first election after the Cambridge Analytica revelations, but also because Brazilian candidates and parties were allowed to devote resources to boost online political content for the first time.[104] A few months before the elections, Facebook announced new measures to improve transparency in political ads and content and address coordinated inauthentic behaviour. These were introduced proactively by the company, also in consideration of the requests made by the Brazilian Electoral Supreme Tribunal (TSE), in a memorandum signed with the company aimed at combating online misinformation.[105]

Facebook also launched the Ad Library and introduced a feature whereby users could easily have access to the information behind ad financing and targeting purposes.[106] Additionally, it attached a specific label to politically sponsored ads. Moreover, Facebook allowed users access to more information concerning Pages, including name alterations and date of creation, which were previously hidden.

---

[104] "Propaganda Eleitoral na Internet" (2018) at 11, online (pdf): *Justiça Eleitoral* https://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet/rybena_pdf?file=https://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet/at_download/file ["Electoral Advertisement on the Internet", translated by the author](although paid political ads remained prohibited, political parties and candidates were allowed to pay to boost their own posts).
[105] Felipe Pontes, "TSE assina memorando com Facebook e Google contra fake news", (28 June 2018), online: *Agência Brasil* <https://agenciabrasil.ebc.com.br/justica/noticia/2018-06/tse-assina-memorando-com-facebook-e-google-contra-fake-news> ["TSE signs memorandum with Facebook and Google against fake news", translated by the author] (where it is however noted that the document does not mention specific actions to be taken but merely states that they should focus on the "prevention of malicious disinformation practices, projects to foster digital education and initiatives that promote quality journalism").
[106] This feature was also being tested during the US 2018 mid-term elections.

These measures were aimed at enhancing users' ability to detect misleading content and were part of the digital literacy campaign led by Facebook in the country.[107] The company also partnered with local media literacy projects and organizations, launched a Fact-Checking program in collaboration with three highly recognized Brazilian fact-checking agencies, developed a marketing campaign and an online de-bunking course, and made available two chat-bots to help users identify the veracity of a post. Although the company confirmed the success of these initiatives, no data about their efficacy has been released, let alone independently verified.

The same can be argued about Facebook's efforts to tackle coordinated inauthentic behaviour and the spread of misinformation. Even though the company claimed that it removed 281 pages and 229 fake accounts between July and October, little is known about their effectiveness.[108] While Facebook praised its removal of fake accounts and misinformation posts, an independent study that analyzed the spread of misinformation on Facebook and YouTube from 2014 to 2020 disagreed. It found that the misinformation links which received the highest engagement in 2019 were officially posted in 2016 and were still present in the platform one year after the elections, thus raising concerns over the fallibility of Facebook's content moderation system.[109] The same study also found that the metrics used by Facebook in portraying the results of its actions were misleading since they only accounted for users'

[107] Katie Harbath, "Protegendo as eleições no Brasil" (24 July 2018), online: *Facebook* <https://about.fb.com/br/news/2018/07/protegendo-as-eleicoes-no-brasil/>. ["Protecting elections in Brazil", translated by the author].
[108] Facebook Inc., "Combatendo a desinformação para proteger a eleição no Brasil - Sobre o Facebook" (23 October 2018), online: *Facebook* <https://about.fb.com/br/news/2018/10/combatendo-a-desinformacao-para-proteger-a-eleicao-no-brasil/> ["Fighting Misinformation to Protect Brazil's Election - About Facebook", translated by the author].
[109] Marco A. Ruedriger & Amaro Grassi, "Desinformação on-line e processos políticos: a circulação de links sobre desconfiança no sistema eleitoral brasileiro no Facebook e no YouTube (2014-2020)" (2020) FGV Policy Paper at 16, online (pdf): <https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/30085/%5bPT%5d%20Estudo%201%20%281%29.pdf?sequence=1&isAllowed=y> ["Online disinformation and political processes: circulation of links about distrust in the Brazilian electoral system on Facebook and YouTube (2014-2020)", translated by the author]

interactions (i.e. comments, shares and likes) with misinformation-related posts, rather than the total views of a post.

According to commentators worldwide, Facebook's efforts seem to have missed the aim once more.[110] The 2018 elections in the world's fourth-largest democracy were characterized by unprecedented levels of misinformation, with a special focus on raising distrust over the electronic ballot system, [111] with Facebook and WhatsApp being the primary means of propagation.[112] After the elections, leaked employee documents revealed that the company failed to identify some of the major misinformation trends and calculate their spread until deletion. Additionally, they showed that most of the content moderation was done ex post upon notification by third parties due to the non-functioning of Facebook ex ante detection tools for the Brazilian context.[113] The disclosure over content moderation practices during the elections remained incomplete and highly subjective. By not setting specific requirements and objectives to be achieved, the vagueness that characterized the memorandum signed between Facebook and the Brazilian regulator allowed the former great discretion in tackling the spread of misinformation on its platforms, which inevitably failed.

---

[110] See Asher Schechter, "Brazil's Election Is Yet Another Indication That Facebook Is Too Big to Manage - ProMarket" (31 October 2018), online: *ProMarket - University of Chicago Booth* <https://promarket.org/2018/10/31/brazils-election-is-yet-another-indication-that-facebook-is-too-big-to-manage/>; Ed Bracho-Polanco, "How Jair Bolsonaro used 'fake news' to win power" (8 January 2019), online: *The Conversation* <https://theconversation.com/how-jair-bolsonaro-used-fake-news-to-win-power-109343>; Tai Nalon, "Did WhatsApp help Bolsonaro win the Brazilian presidency?", *the Washington Post* (November 2018), online: <https://www.washingtonpost.com/news/theworldpost/wp/2018/11/01/whatsapp-2/>.

[111] Mônica Chaves & Adriana Braga, "The agenda of disinformation: 'fake news' and membership categorization analysis in the 2018 Brazilian presidential elections" (2019) 15:3 Brazilian Journalism Research 474 at 486, 492, DOI: <10.25200/BJR.v15n3.2019.1187>.

[112] Tatiana M. S. G. Dourado, *Fake news na eleição presidencial de 2018 no Brasil*, (PhD Thesis, Universidade Federal da Bahia, 2020) [unpublished] at 208. [*Fake news during the 2018 presidential elections in Brasil*, translated by the author] (where the author compares the spread of links containing fake news during the 2018 elections among Facebook, Twitter and WhatsApp); Lucas Vidigal, Gabriela Sarmento & Cida Alves, "Candidatos destinam 1,6% dos gastos da eleição de 2018 para anúncio online, aponta balanço parcial", *G1* (18 September 2018), online: <https://g1.globo.com/politica/eleicoes/2018/noticia/2018/09/18/candidatos-destinam-16-dos-gastos-da-eleicao-de-2018-para-anuncio-online-aponta-balanco-parcial.ghtml>. ["Candidates allocate 1.6% of 2018 election spending to online ads, partial financial statement shows", translated by the author] (where the analysed data shows that Facebook had received 60% of the total political expenditures declared to the TSE reserved to online platform before the official campaign period had started).

[113] Deepa Seetharaman & Jeff Horwitz, "Facebook Touted Its Progress in Brazil Elections. Internally There Were Doubts.", *The Wall Street Journal* (30 August 2019), online: <https://www.wsj.com/articles/facebook-said-it-aced-brazil-elections-internally-there-were-doubts-11567157403>.

3.2. 2019 European Parliament Elections

After the criticism received over its actions in Brazil, Facebook had another chance of demonstrating its improvements in the 2019 European Parliament Elections. Considering the impact of misinformation during previous major political events in the EU, such as the Brexit Referendum in 2016 and the German and French Elections in 2017, the European Commission started strengthening its activities to combat misinformation.

After multi-stakeholder consultations held by the "High-level Expert Group on Fake News and Online Disinformation," the Commission issued its strategy to counter misinformation trends online.[114] Among the Commission's remarks, one is particularly relevant for the scope of this work: *enhanced transparency*. By noting that "[t]he mechanisms that enable the creation, amplification and dissemination of disinformation rely upon a lack of transparency and traceability in the existing platform ecosystem and on the impact of algorithms and online advertising models," the Commission called to action platforms to not only comply with legal requirements but to increase their proactive efforts to fight the phenomenon.[115]

A few months later, the EU Code of Practice on Disinformation was enacted and voluntarily signed by all major platforms, promising to increase their overall actions to tackle online dis- and misinformation, with particular attention to electoral campaigns.[116] The Code also stressed the need to improve transparency over platforms' actions throughout its entirety. Among the requirements, companies were asked to disclose information over political

[114] EU, *Tackling online disinformation: a European Approach*, (Commission COM 236) (26.04.2018), online (pdf): *European Commission* <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&rid=2>.
[115] *Ibid* at 7.
[116] European Union, *Code of Practice on Disinformation*, 2018, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454> [*EU CoP*]. Note that the last major platform to become a signatory party was TikTok in 2020.

advertisement[117] and targeting,[118] collaborate with independent researchers on disinformation,[119] and submit their implementation assessments to third-party review.[120] Additionally, the Code foresaw monthly discussions with its signatories during the first months of its entry into force to allow for an evaluation of its functioning.

To comply with the European regulator's demands, Facebook strengthened its instruments to combat misinformation. It restricted political advertisement by requiring pre-authorization of political ads purchases and made the origin of payment for such ads visible to users through a link redirecting to Facebook's "Ad Library." It also committed to reinforce its efforts against coordinated inauthentic behaviour and expand its fact-checking network.[121] Nevertheless, the changes only became effective one month before the elections, which has led commentators to argue that misinformation had had enough time to spread before the new policy was implemented.[122]

Facebook also promised to remove the content in violation of community standards and to de-prioritize content undermining authenticity. However, it needs to be noted that the company did not provide further details about what content could be considered inauthentic even though in line with community standards. This vagueness further diminishes the efficiency of transparency mechanisms, especially when users whose content has been de-prioritized receive no notification and have no means of contesting the decision. Moreover, no

[117] *EU CoP*, *supra* note 116 at arts 2-4.
[118] *EU CoP*, *supra* note 116 at paras II. B, II. D.
[119] *EU CoP*, *supra* note 116 at arts 12-15.
[120] *EU CoP*, *supra* note 116 at arts 16, 20.
[121] Anika Geisel, "Protecting the European Parliament Elections" (28 January 2019), online: *Facebook* <https://about.fb.com/news/2019/01/european-parliament-elections/>.
[122] Avaaz, *Far Right Networks of Deception* (22/05/2019), online (pdf): *Avaaz* <https://avaazimages.avaaz.org/Avaaz%20Report%20Network%20Deception%2020190522.pdf> at 11 (where it is shown that, prior to take down from Facebook, disinformation network had had 5.9 million interactions daily, with an estimate of 533 million views).

reference is made to data regarding such content in transparency reports, allowing companies like Facebook to act unobserved.[123]

In the compliance report following the European Elections, Facebook provided insights over the results obtained with its actions. It released data concerning the enforcement of its Community Standards, measures addressing coordinated inauthentic behaviour and improvements towards increased independent oversight, such as the award of grants for nineteen research projects studying Facebooks' content moderation policies.[124] However, data provided was particularly general, with little details about the actual effectiveness of such actions. For instance, with regards to the takedown of political ads, the company did not differentiate between ads removed for misleading or false content and ads removed due to their low quality.[125]

Furthermore, one remaining issue with Facebook's attempt to tackle misinformation is that, even though transparency seems to have improved over time, its actual effectiveness has no way of being verified due to independent auditors' lack of access to information.[126] The efforts to improve independent oversight claimed by the company appear doubtful given Facebook's involvement in selecting experts and researchers who are given access to protected

---

[123] Florian Saurwein & Charlotte Spencer-Smith, "Combating Disinformation on Social Media: Multilevel Governance and Distributed Accountability in Europe" (2020) 8:6 Digital Journalism 820 at 832, online: <https://doi.org/10.1080/21670811.2020.1765401>; Maroussia Lévesque, "Applying the Un Guiding Principles on Business and Human Rights to Online Content Moderation" (19 February 2021), Working Paper, DOI: <10.2139/ssrn.3789311> at 16. *Cf* Evelyn Douek, "How Much Power Did Facebook Give Its Oversight Board?" (25 September 2019), online: *Lawfare* <https://www.lawfareblog.com/how-much-power-did-facebook-give-its-oversight-board> (where the author argues that downranking allows Facebook to avoid its Oversight Board to pronounce on controversial cases).

[124] Facebook Inc. *Facebook reports on implementation of the Code of Practice on Disinformation – May report*, (22 May 2019), online (pdf): *European Commission* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60041 at 4-5, 15-16.

[125] European Commission, "Code of Practice on Disinformation - Intermediate Targeted Monitoring - May Reports" (2019) at 3, online: *European Commission* <https://digital-strategy.ec.europa.eu/en/news/last-intermediate-results-eu-code-practice-against-disinformation>.

[126] Florian Saurwein & Charlotte Spenser-Smith, *supra* note 23 at 832. Daniel Boffey, "EU disputes Facebook's claims of progress against fake accounts", *The Guardian* (29 October 2019), online: <https://www.theguardian.com/world/2019/oct/29/europe-accuses-facebook-of-being-slow-to-remove-fake-accounts>.

information.[127] As was the case with the Brazilian elections, once again, doubts over Facebook's claims of success were exacerbated by the opaqueness that characterizes the company's disclosures and by the obstacles encountered by independent researchers and auditors trying to evaluate Facebook's efforts.

3.3. 2020 U.S. Presidential Elections

The latest developments in Facebook's content moderation practices refer to the 2020 U.S. Presidential Elections, which were characterized by an even stronger oversight by scholars and regulators worldwide. Facebook had been under constant pressure to control the influence of problematic behaviour on the electoral results considering the proven Russian manipulation that stained the 2016 elections. In a year where people were spending more time than ever interacting with each other through computer- and smartphone screens, it became essential to limit the spread of online misinformation.

Seeing the results – and failures – of the changes made in Brazil and Europe to address misinformation during the elections, Facebook had to introduce more improvements to its content moderation practices responding to social and regulatory demands for increased efforts to limit the spread of electoral misinformation.

A year prior to the vote, Facebook announced that it was implementing new measures to safeguard the democratic process of the 2020 U.S. Elections. The first changes concerned four main areas: foreign interference, ad- and page-transparency, limits on vote-deterrent advertisements, and misinformation reduction through fact-checking labelling[128]. At first sight, it appears that more significant efforts to counter foreign interference were being devoted to preventing a repetition of the 2016 Russian misinformation campaigns. Indeed, the main

[127] Monika Zalnieriute, *supra* note 48 at 148.
[128] Guy Rosen et al, "Helping to Protect the 2020 US Elections - About Facebook" (21 October 2019), online: *Facebook* <https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/>.

changes announced dealt with improved and more precise identification of page-owners, labelling of state-controlled media, and updated policies on coordinated inauthentic behaviour.

For what concerns misinformation, Facebook announced that it would reduce the distribution of confirmed misinformation from both Facebook and Instagram, de-prioritize content created by accounts regularly spreading misinformation, and apply more explicit labels to posts evaluated by fact-checkers. Additionally, it committed to protecting the integrity of the elections by prohibiting advertisements aimed at discouraging voting and by proactively filtering and removing harmful content through improved machine learning.

Later, Facebook's CEO announced that, after reviewing the efficacy of the implemented changes until then, the company would introduce new limitations on users' engagement with election-related content. More specifically, Zuckerberg explained that Facebook would attach new labels to posts containing misleading information on the outcome of the elections and would provide users with links to authoritative sources such as Reuters and the National Election Pool.[129] Additionally, the company launched the "Voting Information Center," a tool available at the top of users' feeds, providing them with authoritative information about voting registration, procedure, and results. The new measures seemed to indicate a shift in companies' focus from foreign- to domestic interference, especially considering the projected attempts of delegitimizing the outcome of elections after trends emerged on the platform.

One month before the elections, in light of their unprecedented character, Facebook took a step forward and communicated to its users that content aimed at intimidating voters would be banned entirely. It also provided information about the procedures that would be implemented in case of delayed final election results, such as the creation of a new label

[129] Mark Zuckerberg, "The US elections are just two months away …" (3 September 2020), posted on *Mark Zuckerberg*, online: *Facebook* <https://www.facebook.com/zuck/posts/10112270823363411>.

warning users of incorrect and misleading claims.[130] It needs to be noted that the policy mentioned above and procedural changes were widely promoted through user notifications, press statements, and traditional media coverage. However, the same degree of publicity was not found concerning the effectiveness of such containment measures.

In December 2020, Facebook released a report evaluating its actions. While the report repeats in detail the efforts undertaken by the company, as already announced over time on its website, little data concerning the actual results of such measures is provided. For instance, the report mentions that "265,000 pieces of content on Facebook and Instagram in the US were removed between March 1st and Election Day for violating our voter interference policies," but does not mention whether the content has been removed prior to being uploaded due to efficacious automated machine learning screening or whether its removal resulted from ex post review.[131] Moreover, the report states that 180 million posts were labelled as containing misleading information. Still, no data is provided about the effectiveness of such labels besides that only 5% of people click on the post after receiving such a warning.[132]

Furthermore, without access to data by independent scholars and auditors, the veracity of such statements cannot be confirmed. Researchers have complained about platforms' resistance in providing them access to their databases, thus resulting in superficial and incomplete analyses.[133] Such argument is reinforced by a report conducted by the advocacy

---

[130] Guy Rosen, "Preparing for Election Day", (7 October 2020), online: *Facebook* <https://about.fb.com/news/2020/10/preparing-for-election-day/>.

[131] Facebook Inc., *A Look at Facebook and the US 2020 Elections* (December 2020) at 10, online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/12/US-2020-Elections-Report.pdf>. [*FB Elections Report*].

[132] *FB Elections Report*, *supra* note 131 at 11; Rachel Lerman, "Facebook says it labeled 180 million debunked posts ahead of the election", *The Washington Post* (19 November 2020), online: <https://www.washingtonpost.com/technology/2020/11/19/facebook-election-warning-labels/>. It should be remarked that, in the congressional hearing following aftermath of the elections, when asked about the effectiveness of the implemented labelling system, Facebook's CEO did not provide any specific data but promised that it would have been available in the report, which did not happen. See US, *Breaking the News: Censorship, Suppression, and the 2020 Election*: *Hearing before the Committee on the Judiciary*, (17 November 2020) at 04h:12m ff, online (video): *United States Senate Committee on the Judiciary* <https://www.judiciary.senate.gov/meetings/breaking-the-news-censorship-suppression-and-the-2020-election>

[133] Center for an Informed Public et al., *The Long Fuse: Misinformation and the 2020 Election, 2021*, (2021) at 219, 232, online (pdf): *Stanford Digital Repository* <https://purl.stanford.edu/tr171zs0069>; Avaaz, *Facebook:*

group Avaaz, which monitored 100 top-sharers of misinformation on Facebook and found that the company only effectively addressed misinformation from October 2020, when the interactions with such pages suddenly dropped.[134] The study shows that Facebook's report "appears to be biased, particularly since the platform does not highlight what it could have done, it does not highlight how many people saw the misinformation before it had acted on it, and does not provide any measure for the harms caused by a lack of early action."[135]

4. Improving transparency-reliant systems through standardization and key-performance-indicators

After having identified the shortcomings of transparency mechanisms and the recurring failures of companies' improvements, it is now possible to see how these could be overcome. It is necessary to remark that both regulators and the industry itself could apply the following proposals to improve the overall system of transparency of content moderation practices. Moreover, the suggestions made are thought to remain flexible to be easily adaptable to the type of platform and the technological developments of social media. Nevertheless, it is important to note that, to achieve such aims, the shift needs to be widespread across the industry and not be limited to a few platforms in an enclosed jurisdiction.

A common characteristic among the shortcomings identified in Section 2 was the lack of standardization among transparency practices of social media platforms. Indeed, when platforms use different metrics in their transparency reports, different types of user

From Election to Insurrection - How Facebook Failed Voters and Nearly Set Democracy Aflame (18 March 2021) at 5, online: *Avaaz* https://secure.avaaz.org/campaign/en/facebook_election_insurrection/; Irene Pasquetto & Briony Swire-Thompson, "Tackling misinformation: What researchers could do with social media data", (9 December 2020) [Avaaz (b)], online: *Harvard Kennedy School Misinformation Review* <https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/>.

[134] *Avaaz (b)*, *supra* note 133 at 6.

[135] *Avaaz (b)*, *supra* note 133 at 8 (where it is observed that the delay resulted in over 10.1 billion estimated views of posts containing disinformation that could have been avoided, had the company acted earlier).

notifications, and different templates in compliance reports, it becomes particularly difficult to cross-compare platforms and evaluate their performance. This de-alignment among transparency practices results in a distortion of the information available, which inevitably misses the aim of improving the efficacy of content moderation.[136] Moreover, this heterogeneity makes it highly complex for researchers and experts to identify current online misinformation trends and determine where more resources are needed. This is why the standardization of transparency mechanisms is the first step in improving the efficiency of content-moderation practices.[137] By standardization, it is meant that platform transparency practices should adopt the same minimum degree of specificity over the data disclosed.[138]

Standardization has already proved its efficacy in other areas of the law. An example of its use can be found in the success of the Global Reporting Initiative (GRI), which is the leading transparency mechanism used in corporate, social, and environmental responsibility.[139] Based on three main pillars – multistakeholder participation, institutionalization, and stewardship – the organization has improved sustainability transparency by developing 39 different sets of ESG standards which are now the most used by companies worldwide.[140]

---

[136] See generally Christopher Parsons, "The (In)effectiveness of Voluntarily Produced Transparency Reports" (2017) 58:1 Business & Society 103 at 106, online: <https://journals.sagepub.com/doi/full/10.1177/0007650317717957> (discussing corporate social responsibility and transparency reports).

[137] Standardization has already been proposed as a fundamental aspect of improved targeted transparency. See Fung et al., *supra* note 53 at 28 ff.; Jennifer C. Daskal, "Speech Across Borders" (2019) 105:8 Virginia Law Review 1605 at 1664, online: *Virginia Law Review* <https://www.virginialawreview.org/articles/speech-across-borders/>.

[138] Fung et al., *supra* note 53 at 43.

[139] See generally Halina Szejnwald Brown, Martin de Jong & Teodorina Lessidrenska, "The rise of the Global Reporting Initiative: a case of institutional entrepreneurship" (2009) 18:2 Environmental Politics 182, DOI: <10.1080/09644010802682551>; Mikkel Flyverbom, Lars Thøger Christensen & Hans Krause Hansen, "The Transparency–Power Nexus" (2015) 29:3 Management Communication Quarterly 385 at 401, online: <https://journals.sagepub.com/doi/10.1177/0893318915593116 >.

[140] Halina Szejnwald Brown, Martin de Jong & Teodorina Lessidrenska, supra note 139 at 190-193; KPMG, *The KPMG Survey of Corporate Responsibility Reporting 2017*, by Jose Luis Blasco & Adrian King (2017), online (pdf): *KPMG* https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/10/kpmg-survey-of-corporate-responsibility-reporting-2017.pdf ("[t]he majority of N100 (74 percent) and G250 companies (89 percent) are using some kind of guidance or framework for their reporting. The GRI framework is the most commonly used, with 63 percent of N100 reports and 75 percent of G250 reports applying it.")

Some governmental and industry-wide initiatives have also recognized the efficacy of standardization in the fight against misinformation. For example, the European Commission has defined standardization as fundamental to better the impact of the EU Code of Practice on Disinformation after commentators have complained about the inefficacy of discrepancies across platforms' reporting.[141] Among the industry, standardization has also been the focus of the Information Trust Alliance, an initiative aimed at addressing misinformation in India, where social media companies have shown their commitment to establishing a standardized approach to address misinformation in a draft Code of Practice.[142]

While increasing standardization of transparency mechanisms is a fundamental step to improve the legibility of a company's disclosed data, establishing common key-performance indicators (KPIs) is equally necessary to facilitate the evaluation of platforms' actions and performance over time, as well as the spread of online misinformation among them.[143] Indeed, KPIs serve to establish the effectiveness of a company's achievements in light of a set of standards.

Their use has already been discussed among national and supranational regulators, who have argued that the differences among platforms should not prevent the establishment of common assessment standards in platforms' achievement of policy aims.[144] In other sectors –

---

[141] EU, *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*, (Commission SWD 180) (10 September 2020) at 23, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212>.

[142] Megha Mandavia, "Social media to join hands to fight fake news, hate speech", *The Economic Times* (19 February 2020), online: <https://economictimes.indiatimes.com/tech/internet/social-media-to-join-hands-to-fight-fake-news-hate-speech/articleshow/74200542.cms?from=mdr>. However, such efforts have been repeatedly pushed back by a lack of consensus among participants.

[143] Ben Wagner & Carolina Ferro, *supra* note 87 at 19; European Commission, "Study for the Assessment of the Implementation of the Code of Practice on Disinformation", by Iva Plasilova et al. (May 2020) at 89, online: *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=66649> [*VVA Report*]

[144] EU, *Guidance on Strengthening the Code of Practice on Disinformation*, (Commission COM 262) (26 May 2021) at 21-22, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/redirection/document/76495> (according to the EU Commission, these KPIs should be divided in two groups – "service-level indicators" and "structural indicators" – that assess both the policies introduced by platforms and the impact of the requirements imposed by the regulator); Australian Communications and Media Authority, *Misinformation and news quality on digital platforms in Australia* (June 2020) at 32-33, online (pdf): *ACMA* <https://www.acma.gov.au/sites/default/files/2020-06/Misinformation%20and%20news%20quality%20position%20paper.pdf>.

such as finance and insurance, IT-services, manufacture, and healthcare – KPIs have already been successfully implemented and proved their efficiency.[145] An example of such can be found in the adoption of KPIs in financial reporting, which has been reinforced by several national and international legislation and recommendations to increase clarity and comparability among companies' disclosures, especially when such companies operate in multiple jurisdictions.[146] Similarly, adopting shared industry KPIs in platforms' reporting could increase clarity and comparability among them, with a view to hold platforms to their claims and encourage progress over time. The following subsections will propose four areas for standardization of transparency mechanisms and their respective potential KPIs.[147]

4.1. Ex Ante and Ex Post Review

Given the scale of the content moderation operations that major social media platforms are confronted with at every second, all platforms have heavily invested in automated content moderation systems that screen content before its upload. Despite the significant improvements in the past years, at present, algorithms are not always able to identify misinformation correctly, which generates both false positives and false negatives. This is why most platforms provide users with the option of appealing a decision while also relying on human moderators and third-party fact-checkers to look for and assess ambiguous content ex post.

However, not all platforms specify whether content has been removed by an algorithmic detection tool or whether it has been flagged and identified as misinformation after it was already circulating on the platform. This lack of specificity hinders an assessment of the

[145] See "Key Performance Indicators Listed by Sector" (2016), online (pdf): *IntraFocus* <https://static.intrafocus.com/uploads/2016/02/Key-Performance-Indicators-by-Sector.pdf>

[146] See e.g. EU, *European Parliament and Council Directive 2013/34/EU of 26 June 2013 on the annual financial statements, consolidated financial statements and related reports of certain types of undertakings, amending Directive 2006/43/EC of the European Parliament and of the Council and repealing Council Directives 78/660/EEC and 83/349/EEC*, [2013] OJ, L 138/19 at 3.

[147] The suggested KPIs were drafted starting from the suggestions made by Iva Plasilova et al. in "European Commission, "Study for the Assessment of the Implementation of the Code of Practice on Disinformation", (May 2020) at 89-95, online: *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=66649>".

overall spread of the phenomenon. Indeed, this distinction is particularly relevant since it provides data useful to determine the reach that such content has received. If the content is removed during an ex ante review, its reach is zero. In contrast, if the content is only detected after circulation, its reach increases exponentially by the minute, which requires more intense countering mechanisms, such as increasing the visibility of correct information debunking it and notifying users in a timely manner. Therefore, platforms should differentiate misinformation identified ex ante and ex post in their reporting and disclose whether such identification was made by algorithmic screening or by a human moderator. Providing this information in Facebook's transparency report over the US 2020 Elections would have contributed to a better evaluation of the spread of electoral misinformation in the country at the time.[148]

Moreover, given the fallibility of algorithmic detection, platforms should also disclose the number of instances of algorithmic false positives[149] and false negatives.[150] This information would allow observers to assess the dynamics between algorithms and human moderators to tackle misinformation, thus providing a clearer picture of social media companies' overall content moderation operations.[151] Doing so could bring to light possible deficiencies in platforms' practices, such as the failure of Facebook ex ante detection systems in Brazil during the 2018 Elections.[152]

Platforms could present this information by referring to the KPI of the *ratio between the false positives and the total number of content removed via algorithmic screening*. Such disclosure is necessary to enable the evaluation of the algorithmic effectiveness of platforms' screening systems. Through a cross-comparison of the data provided, a margin of algorithmic

[148] See Section 3.3, *above*.
[149] I.e., when a user has appealed an automated decision and a human moderator has confirmed the appeal.
[150] I.e., when disinformation is identified through users' and human moderators' flagging.
[151] Petros Iosifidis & Nicholas Nicoli, "The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation" (2019) 82:1 International Communication Gazette 60 at 74.
[152] Deepa Seetharaman & Jeff Horwitz, *supra* note 113 at 112.

error could be established. This would be beneficial both for platforms' reputation among its users and regulators willing to impose more stringent oversight over platforms' declarations. Indeed, platforms have constantly been advertising the technological advancements being made on their moderation algorithms, but there is no way to verify such improvements. This issue could be overcome with a disclosure of such margin of error, which would provide regulators with a means to evaluate platforms' improvements over time.

Overall separating data over ex ante algorithmic screening and ex post review would, on the one hand, increase the information available over the current state-of-the-art of screening algorithms, setting realistic expectations on both users and regulators. On the other, it would facilitate the assessment of the capacity of human moderators and eventually signal the need for improvements in terms of the numbers of individuals tasked with content moderation functions and training received.

4.2. Degree Of Decision-Making

Platforms should disclose the means and degrees of decision-making involved in content moderation practices. In addition to clearly distinguishing between how much content has been automatically reviewed and removed by the algorithms and how much has been reviewed by human moderators, platforms should improve transparency over the degree of human decision-making involved in a moderation decision.[153] Given the complexity in addressing misinformation that characterizes some decisions,[154] it is relevant to know whether it has been made by a contractor, platform employee, or upper-level executive. This could be done by differentiating data disclosed according to the three degrees of decision-making. Such

[153] Timothy Ash, Robert Gorwa & Danaë Metaxa, *supra* note 93 at 12.
[154] Akriti Gaur, "Towards Policy and Regulatory Approaches for Combating Misinformation in India" in Michael Karanicolas, ed, *Tackling the "Fake" Without Harming the "News": A Paper Series on Regulatory Responses to Misinformation* (Wikimidia/Yale Law School, 2021) 30 at 46-47; Peter Cunliffe-Jones, "Europe's latest export: A bad disinformation strategy", *Politico* (7 June 2021), online: <https://www.politico.eu/article/europe-bad-disinformation-strategy-digital-services-act-dsa/amp/>.

information would allow companies and external observers to identify the necessary resources and improvements needed in content moderation protocols.

When misinformation is supported by highly influential figures (e.g. political leaders), its virality increases exponentially as well as its impact on society.[155] However, deciding whether to remove content posted by a political leader or an equally popular speaker is particularly complex given the possible conflicts with other fundamental rights such as freedom of speech and political expression. Moreover, commentators have been divided when discussing the ban from social media access or the deletion of content posted by national leaders. Those against it argue that it might endanger the freedom of speech of political actors.[156] In contrast, others in favour of removing misleading content and misinformation believe that policies should be applied consistently, regardless of their author.[157] Given this complexity, such decisions are often escalated (i.e., to pass on the decision-making) to upper-level executives. Knowing the degree and impact of executive decision-making is necessary to identify not only the gaps in platforms' content moderation practices but also eventual

[155] Soroush Vosoughi, Deb Roy & Sinan Aral, "The spread of true and false news online" (2018) 359:6380 Science 1146–1151, DOI: 10.1126/science.aap9559 ("[f]alsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information").

[156] See e.g. "Merkel kritisiert Twitter-Sperre für Trump",*Tagesspiegel* (11 January 2021), online: <https://www.tagesspiegel.de/politik/meinungsfreiheit-von-elementarer-bedeutung-merkel-kritisiert-twitter-sperre-fuer-trump/26786886.html> [translated by the author] (where the spokesman of the German Chancellor Angela Merkel commented that "the basic right to freedom of expression is of fundamental importance," said Seibert. Interventions can therefore only take place in accordance with the law and within the framework defined by the legislator and "not after the decision of the corporate management of social media platforms"); Alexej Navalny, "1. I think that the ban of Donald Trump on Twitter is an unacceptable act of censorship (THREAD)" (9 January 2021), online: Twitter https://twitter.com/navalny/status/1347969772177264644?s=20 (where, in a long Twitter thread, the Russian opposition leader said that "This precedent will be exploited by enemies of free speech worldwide. Also in Russia").

[157] The debate has been particularly rich in concurring opinions after the ban of President Trump from major social media after the Storming of Capitol Hill in January 2021. See e.g. Anti-Defamation League et al., "Facebook's Suspension of Donald J. Trump" (12 February 2021), online (pdf): *Free Press* (where in a letter addressed to the Facebook Oversight Board the authors urged for the permanent ban of President Trump from social media due to "act in the public interest and prioritize the health and safety of our communities"); Center for Democracy and Technology, "Comments to Facebook Oversight Board" (11 February 2021), online (pdf): *Center for Democracy and Technology* <https://cdt.org/wp-content/uploads/2021/02/CDT-comments-to-FB-Oversight-Board-on-2021-001-FB-FBR.pdf> ("In view of the potential violence and physical injury that speech from political leaders can incite, account suspensions can be an appropriate enforcement action").

decisional biases. Tackling online misinformation requires that policies and their enforcement be non-discretionary, clear, well-defined, and easily understandable by its users. Therefore, communicating how decisions with significant societal impacts are made is essential primarily to ensure their consistency and coherence over time and across the entire platform.

To enable meaningful comparison among companies, they should be evaluated in light of two KPIs. The first has already been suggested by the VVA Report in the EU, according to which platforms should disclose the *ratio of directly contracted employees tasked with identifying and deactivating disinformatio*n.[158] Disclosing such information would allow for a comprehensive comparison of the financial resources invested by companies in content moderation practices. This disclosure would then be used as a reference for regulators when drafting legislation demanding more action from platforms. Furthermore, a comparison over time of financial resources devoted to content moderation could be helpful when assessing companies' allocation of resources to limit the spread of misinformation on their platform.

The second KPI that should be used to measure the impact and performance of platforms' actions to tackle online misinformation is the *ratio of executive decision-making per country*. Disclosing such data would provide useful information to both platforms and regulators over the geographical limitations and deficiencies of companies' practices aimed at countering the spread of misinformation. Moreover, such data could support (or dismantle) claims of political bias in platforms' executive decision-making.[159] Lastly, producing such data would be useful to identify companies' priorities and lack thereof with regard to specific

[158] *VVA Report*, *supra* note 143 at 89.
[159] See Chinmayi Arun, "Facebook's Faces" (2021) at 3,12, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3805210> (where the author argues that "Facebook does not engage with all states in the same way, and it certainly does not engage with all publics in the same way" and criticizes the company for responding to "summons by the European Parliament but refus[ing] calls from other Parliaments around the world including Australia and India, and even refused to appear before an 'international grand committee' consisting of policy makers from Argentina, Brazil, Canada, Ireland, Latvia, Singapore and the United Kingdom").

geographic areas,[160] which could thus incentivize increased regulatory action and company's practices reform where needed.

4.3. Engagement with Misinformation

The phenomenon of online misinformation unfolds in different ways that have been defined incoherently among and within companies through reactive ex post measures.[161] However, this incoherence should not be used to obstacle the analysis of the spread of misinformation. Although some regulatory instruments have already started to establish a common framework for defining the phenomenon,[162] the terminology is still used incoherently across platforms, regulators, and academia. The categories of disclosure regarding misinformation should be progressively uniformed amongst platforms and regulators. This would avoid the current situation where terms are used inconsistently even across the same platform, given that these inconsistencies are then reflected in transparency mechanisms. I suggest three categories to be differentiated under the broader umbrella of "misinformation" –

[160] Social media companies have been criticized for the discrepancies in addressing misinformation and content manipulation between the Global North and the Global South, the latter being often considered as "non-priority" and politically motivated. See especially Craig Silverman, Ryan Mac & Pranav Dixit, "Whistleblower Says Facebook Ignored Global Political Manipulation", (14 September 2020), online: BuzzFeed News <https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo>. *Cf* Newley Purnell & Jeff Horwitz, "Facebook's Hate-Speech Rules Collide With Indian Politics", Wall Street Journal (14 August 2020), online: <https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346> (where it is reported that Facebook has refused to apply its hate speech policy to Indian politicians since it "would damage the company's business prospects in the country").

[161] The same happened with the denominations of Corporate Social Responsibility reporting, where companies developed different denominations over time, which was then replaced with a harmonized approach by the Global Reporting Initiative in the early 2000s. See Ondina Gabrovec Mei, "CSR and Social Reporting: Moving Towards Standardization" (2013) [unpublished], online (pdf): *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2239658&download=yes. The terms most used by platforms vary among: "information operations", "computational propaganda", "information manipulation", "information warfare", "information disorder", "hybrid warfare", "strategic deception", "manipulative interference", "inauthentic activities", "malicious automation", "coordinated inauthentic behavior", and "misrepresentation". See James Pamment, The EU Code of Practice on Disinformation: Briefing Note for the New EU Commission (Carnegie Endowment for International Peace, 2020) at 3-4.

[162] EU, *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*, (Commission SWD 180) (10 September 2020), online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212> at 12-13.

"information operations," "manipulated media," and "inauthentic behaviour"[163] – considering the different countering mechanisms they require. Indeed, the penalties imposed by platforms in case of breach of community standards are a crucial element in the fight against online misinformation. These can go from diminishing the visibility of the problematic content to removing the overall access of the spreader to the platform. Therefore, given that platforms do not disclose data over all countering mechanisms available, differentiating the three seems to be fundamental to improve transparency.

Firstly, implementing uniform categories of disclosure would enable a more efficient and meaningful cross-comparison of platforms' actions and transparency reports by improving the clarity and legibility of disclosed data.[164] This would also draw attention to problematic trends and drive further targeted improvements. Secondly, it could encourage the adoption of consistent terminology over time in regulatory initiatives at the national and supranational levels. Creating a shared agreement over the scope and aims to be achieved and providing platforms with clear definitions and requirements at a global scale is necessary given the worldwide reach of internet platforms.[165] Thirdly, it would promote collaboration among platforms to ensure that misinformation-related behaviour is hindered and discouraged in its totality,[166] thus avoiding the situation where actors migrate to platforms with fewer and less specific restrictions.[167]

---

[163] To ensure that the categories remain flexible enough to encompass future developments and all related behaviour, broader terminology should be preferred to address the widespread phenomenon of disinformation. At the same time, the use of specific terms should be limited to sub-categories. See EU, *Guidance on Strengthening the Code of Practice on Disinformation*, (Commission COM 262) (26 May 2021) at 12, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/redirection/document/76495> [*EU COM (a)*]. Nevertheless, given the fast-changing pace of technology, the categories should be periodically reviewed in a collaborative setting, including all stakeholders, to ensure that new issues are correctly identified and reacted to in a consistent way across platforms, regulators, academics, and civil society.

[164] James Pamment, *supra* note 161 at 4.

[165] *Ibid*. This has already been suggested in the context of the EU Code of Practice, after companies justified the gaps in their transparency reports by complaining that the requirements of the Code were not specific enough.

[166] EU, *Guidance on Strengthening the Code of Practice on Disinformation*, (Commission COM 262) (26 May 2021) at 12, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/redirection/document/76495> [*EU COM (a)*].

[167] See ref. to Parler case and Capitol Hill Violence in Section 2.

"Information operations" is a phrase originally used within military organizations to define technological operations aimed at negatively influencing, disrupting, and corrupting the adversary targets. In the field of social media, the term has also been used to refer to the dissemination of misleading information aimed at manipulating and influencing the audience to interfere with democratic processes. Some platforms already mention information operations either within their policies or in their transparency reports.[168] However, at present, the data provided by platforms has been restricted to foreign state-led initiatives trying to interfere with adversary countries. This limitation is also reflected in the countermeasures adopted. Nonetheless, recent independent reports have highlighted increasing trends in domestic information operations. Therefore, both domestic and foreign initiatives should be taken into consideration by platforms when calculating data to be provided in transparency reports referring to information operations on social media.

The second aspect of misinformation that should be included in transparency reports is "manipulated media." [169] The phenomenon is described in relatively similar ways among

[168] See e.g. Facebook Inc., *Information Operations and Facebook*, by Jen Weedon, William Nuland & Alex Stamos (27 April 2017), online (pdf): < https://about.fb.com/wp-content/uploads/2017/04/facebook-and-information-operations-v1.pdf > (Facebook defines it as "actions taken by organized actors (governments or non-state actors) to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome" at 5); Twitter Inc., "Information Operations Report" (2021), online: *Twitter* <https://transparency.twitter.com/en/reports/information-operations.html> (Twitter describes it as "Platform manipulation that we can reliably attribute to a government or state linked actor is considered an information operation").

[169] While the term "fake news" has become popular among politicians to refer to the phenomenon of misinformation, scholars agree that it is not appropriate to encompass the entirety of the phenomenon. See e.g. Bente Kalsnes, "Fake News" (2018) Oxford Research Encyclopedia of Communication, online: *Oxford Research Encyclopedia* <https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-809> ("politicians and other powerful actors have appropriated the term to characterize media coverage they do not like"); Data & Society, Media manipulation and disinformation online, by Alice Marwick & Rebecca Lewis (2017), online (pdf): Data & Society <https://datasociety.net/wp-content/uploads/2017/05/DataAndSociety_MediaManipulationAndDisinformationOnline-1.pdf> ("the term itself has quickly become contentious and politically-motivated" at 44); Steven Erlanger, "'Fake News,' Trump's Obsession, Is Now a Cudgel for Strongmen (Published 2017)", *The New York Times* (12 December 2017), online: <https://www.nytimes.com/2017/12/12/world/europe/trump-fake-news-dictators.html> ("[a]round the world, authoritarians, populists and other political leaders have seized on [it] as a tool for attacking their critics and, in some cases, deliberately undermining the institutions of democracy").

platforms as "media altered to mislead and manipulate others"[170] and has been the focus of discussions among civil society,[171] leading scholars, and international organizations worldwide.[172] Considering manipulated media by itself is necessary since it would imply the disclosure of data of de-prioritized content. Indeed, the penalties imposed by platforms in case of breach of community standards are a crucial element in the fight against online misinformation. These can go from diminishing the visibility of the problematic content to removing the overall access of the spreader to the platform. Recognizing the difficulties in identifying manipulated media, the platforms' first approach to potential problematic content is often to diminish its visibility on users' feeds and algorithmic recommendations, i.e. "de-prioritization." This de-prioritization often happens when another user or fact-checkers flag a piece of content or when the post's patterns of engagement are identified as suspicious by the platform's algorithm.[173]

While companies admit the use of de-prioritization tools, related data is often lacking from transparency reports, which leaves a gap and provides an incomplete picture of platforms' actions to counter misinformation. Commentators had already remarked this issue in the aftermath of the European Parliament Elections in 2019, when Facebook did not release any data over de-prioritized content.[174] However, platforms are still reluctant to disclose such information. Given that manipulated media is often de-prioritized rather than

---

[170] See e.g. *FB CS*, *supra* note 64 at "Manipulated Media" (where Facebook defines it as "media where the manipulation is not apparent and could mislead, particularly in the case of video content"); *Twitter Rules*, *supra* note 65 at "Manipulated Media" (where Twitter explains it as "media (videos, audio, and images) that have been deceptively altered or fabricated in ways that mislead or deceive people about the media's authenticity where threats to physical safety or other serious harm may result); and Google LLC., "Spam, deceptive practices, & scams policies - YouTube Help" (2021),online: *Google* <https://support.google.com/youtube/answer/2801973?hl=en> (where YouTube clarifies that it relates to "content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm").

[171] See e.g. Alice Marwick & Rebecca Lewis, *supra* note 169 at 2-4.

[172] See e.g. Yasha Lange, *Media and Elections* (Council of Europe Publishing, 1999), online (pdf): <https://rm.coe.int/0900001680483b46>.

[173] Data & Society, *Dead Reckoning. Navigating Content Moderation After "Fake News"*, by Robyn Caplan, Lauren Hanson & Joan Donovan (February 2018) at 21, online (pdf): *Data & Society* <https://datasociety.net/pubs/oh/DataAndSociety_Dead_Reckoning_2018.pdf>.

[174] See Section 3.2, *above.*

removed, including it as a specific category would also encourage companies to broaden the overall disclosure of platforms' policy-enforcement.

Lastly, the third aspect deals with users who use social media in ways that platforms define as *inauthentic*, that is, that engage in misrepresentation, impersonation, and artificial manipulation of interactions with other accounts. While the consequences of one individual user misrepresenting themselves might be more limited (although not devoid thereof), a network of inauthentic accounts acting together towards a collective aim can have a much more significant impact on society.[175] These kinds of actions – so-called "coordinated inauthentic behavior" have been recognized by scholars as a leading strategy for malicious actors who employ users to spread misleading content and impersonate others, with the goal of manipulating public discourse.[176] Inauthentic behaviour should be understood as account networks interacting with each other in a deceptive way, including those who receive payment for such interactions.[177] In light of this, all major platforms should disclose data over the presence of inauthentic behaviour – both in terms of individual and coordinated actions – separately. This will provide observers with specific data over the recurrence of the phenomenon and the effectiveness of platforms in countering it.

For each category mentioned above, referring to the engagement each one receives could be relevant to measure their overall reach and impact. However, to avoid misleading disclosures by companies when calculating the efficacy of their detection and countering

[175] Fabio Giglietto et al, "It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections" (2020) 23:6 Information, Communication & Society 867, DOI: <10.1080/1369118X.2020.1739732>.
[176] Fabio Giglietto et al., *supra* note 175 at 872. See e.g. Tobias R Keller et al., *#ARSONEMERGENCY: Climate Change Disinformation During the Australian Bushfire Season 2019-2020* (paper delivered at the 21st Annual Conference of the Association of Internet Researchers, 27-31 October 2020), [unpublished] (on the spread of misinformation through coordinated inauthentic behaviour in Australia); Patrícia Rossini et al., "Explaining Dysfunctional Information Sharing On Whatsapp And Facebook In Brazil" (paper delivered at the 21st Annual Conference of the Association of Internet Researchers, 27-31 October 2020), [unpublished].
[177] This broader definition has been developed by the EU Disinformation Lab. See Antoine Grégoire, "CIB Detection Tree: 2nd Branch" (14 June 2021), online: *EU DisinfoLab* <https://www.disinfo.eu/publications/cib-detection-tree2/>.

systems, such as happened with Facebook during the Brazilian Elections in 2018, a distinction should be made between passive and active engagement. That is, engagement comprising all the time the content has appeared in a user's feed, and engagement considering only the number of interactions such content has received (e.g. likes, shares, comments, etc.).[178] Given the different consequences that passive and active engagement has on users, understanding both kinds is necessary when evaluating the impact of each phenomenon and the efficacy of the countering mechanisms adopted by social media platforms.[179]

Calculating the difference between the active engagement of misinformation and that of other posts serves to understand the dynamics with which misinformation is spread across platforms. To do so, the *ratio between the active engagement misinformation-related content receives prior removal and the active engagement other content receives* could be used by platforms as a KPI. Disclosing this difference would enable observers to establish what makes different kinds of misinformation thrive.[180] Given that each category requires different responses, this information is fundamental to develop efficient countering mechanisms.[181]

---

[178] See Philippe Verduyn et al, "Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review" (2017) 11:1 Social Issues and Policy Review 274 at 281, DOI: <10.1111/sipr.12033> (according to which, "[a]ctive usage refers to activities that facilitate direct exchanges with other(s)" whereas "[d]uring passive usage of social network sites, information is typically consumed without communicating with the owner of the content).

[179] Briana M. Trifiro & Jennifer Gerson, "Social Media Usage Patterns: Research Note Regarding the Lack of Universal Validated Measures for Active and Passive Use" (2019) 5:2 Social Media + Society, DOI: <10.1177/2056305119848743> ("[t]he ability to compare and understand how […] different levels of engagement with those platforms impact users can help identify which aspects of social media use are beneficial, and which have a negative impact on its users across platforms").

[180] Spreaders of misinformation tend to have the highest levels of active engagement. See Kevin Roose, "Inside Facebook's Data Wars", The New York Times (14 July 2021), online: <https://www.nytimes.com/2021/07/14/technology/facebook-data.html>; Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement" (5 May 2021), posted on *Mark Zuckerberg*, online: *Facebook* https://www.facebook.com/notes/751449002072082/ ("when left unchecked, people will engage disproportionately with more sensationalist and provocative content").

[181] Online Civic Culture Centre, *News Sharing On UK Social Media: Misinformation, Disinformation, And Correction*, by Andrew Chadwick & Cristian Vaccari (2019), online: *Loughborough University* < https://www.lboro.ac.uk/research/online-civic-culture-centre/news-events/articles/o3c-1-survey-report-news-sharing-misinformation/> ("[e]xploring why, and with what effects, people share news about politics on social media is therefore an essential part of the broader debate about the relationship between the internet and democracy"); *VVA Report*, *supra* note 143 at 94.

On the other hand, the *ratio of total passive engagement that misinformation content receives compared to other content* would allow for a better assessment of social media recommender systems. When the engagement of problematic content is greater, it could be a signal that algorithms are prioritizing it. Considering that recommender systems are designed to prolong users' permanence on the platform, identifying whether users are being encouraged to come across misinformation is fundamental since it would immediately pressure platforms to modify their design.[182] Furthermore, this could encourage platforms to exploit the power of these algorithms and enhance engagement of authoritative content to counter misinformation trends.[183]

4.4. User Awareness

To tackle online misinformation also means to improve users' experience when navigating social media platforms. Users engage with content in different ways, ranging from passively scrolling through their feeds to reading, watching, commenting, and sharing content created by others. When users interact with misinformation received by friends or family members, such information tends to be perceived as true in a more unconscious way than information obtained via other means.[184] Studies have also demonstrated that users are more likely to re-share content when it is initially shared by others in their social circle.[185] Yet, it has

[182] For instance, after a Facebook's internal report showing that 64% of the members in extremist groups have joined upon Facebook's own recommendation tool was leaked, the company decided to shut down algorithmic group recommendations. See Jeff Horwitz & Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", The Wall Street Journal (26 May 2020), online: <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>.

[183] Maria A Golino, "Algorithms in Social Media Platforms" (24 April 2021), online: *Institute for Internet & Just Society* <https://www.internetjustsociety.org/algorithms-in-social-media-platforms> ("algorithms may be created with the aim of increasing awareness […] on a specific matter, some users may suddenly see in their feed an increase of posts concerning [a specific topic]").

[184] Lavinia Marin, "Sharing (mis) information on social networking sites. An exploration of the norms for distributing content authored by others" (2021) Ethics and Information Technology, DOI: 0.1007/s10676-021-09578-y ("since sharing amplifies misinformation to an unprecedented extent, it generates epistemic harms at collective and individual levels. The individual harm is that some people may acquire misleading beliefs as result of seeing misinformation shared by their peers" at 1).

[185] Wendy W Moe & David A Schweidel, "Why Do We Share Our Opinions?" in *Social Media Intelligence* (Cambridge University Press, 2014) 37; Elle Hunt, "What is fake news? How to spot it and what you can do to

also been proved that users care about their reputation among friends and family and do not share information they recognize as false. Therefore, alerting users is fundamental to limit the spread of misinformation.

User awareness policies have so far been focused on media literacy campaigns which aim at empowering users to autonomously recognize misinformation.[186] These are thought to fill the gap left by imperfect algorithms and limited fact-checking systems. Platforms have also developed tools available to users re-directing them to authoritative sources such as information centres and explanatory labels attached to problematic content. However, little about the efficacy of these measures is disclosed. For instance, when Facebook praised the success of its media literacy campaign during the Brazilian elections in 2018, no data was available to support such claims.[187] This lack of disclosure hinders the assessment of each tools' performance.

To improve transparency in this regard, the *ratio between users who have interacted with tools designed to fight misinformation against all instances where such tools were available* should be disclosed by companies and used as a KPI.[188] Indeed, this information would avoid circumstances such as that of Brazil 2018 and that of the US 2020 Elections, where Facebook was not able to provide data over the effectiveness of the labels it applied to electoral content.[189]

---

stop it", *The Guardian* (17 December 2016), online: <https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>;
Jan H. Kietzmann et al, "Social media? Get serious! Understanding the functional building blocks of social media" (2011) 54:3 Business Horizons 241–251, online: <https://doi.org/10.1016/j.bushor.2011.01.005>.

[186] User awareness has already been embraced by some regulators in proposals for platform governance. For instance, in the EU Code of Practice, improving users' empowerment is one of the key-requirements to signatories. See *EU CofP, supra* note 116 at 6-7.

[187] See Section 3.1, *above.*

[188] This KPI has been suggested in the *VVA Report*, *supra* note 143 at 94.

[189] Similarly, Twitter was also unable to quantify the efficacy of labels attached to tweets. *Cf* US, *supra* note 132 at 1:56.

Furthermore, another user empowerment tool is particularly relevant in fighting misinformation: user notifications. That is, directly informing users who have previously interacted with content later identified as misinformation. This practice is not yet widespread across platforms, even though studies have remarked their usefulness. To encourage a more extensive adoption of user notification, they should be explicitly included among the different user empowerment tools. This could be reinforced by adopting the *ratio between the number of users notified after having interacted with misinformation and its overall engagement* as a KPI. By requiring that platforms disclose data regarding the operation of their user empowerment tools, companies could thus be encouraged to make user notifications the default rather than the exception if numbers appear favourable.

Referring to both these KPIs in transparency reports would allow observers to measure the performance of user empowerment tools and whether they are able to attract users' attention and counter the effects of misinformation. This assessment could potentially lead platforms to strengthen their action by introducing new features and rethinking the least-performing ones, through the public pressure and oversight generated by disclosure. At the same time, such information could also serve as guidance to government-led initiatives aimed at empowering users, such as media literacy programs and awareness campaigns.[190]

5. Conclusion

Recent events have shown that no democracy is safe from the dangers of misinformation. If social media have eased communication amongst individuals and improved access to multiple sources of information, they have also been the ace in the hole of malicious

[190] Media literacy programs such as those developed by Facebook in partnership with the Brazilian government during the electoral campaign on 2018.

actors trying to disrupt societies worldwide. From conspiracy theorists to populist leaders and adversary governments, social media have proved efficient in spreading one's message to an unprecedented audience in an extremely short amount of time.

I argue that transparency has the potential of strengthening both platforms and regulators' actions to counter the spread of misinformation online. It does so by enhancing the understanding of how it spreads, how quick platforms are in identifying it, in which geographical areas it is most common, and how users interact with it. However, I show that current transparency mechanisms fail to deliver the expected outcomes since they present significant shortcomings due to their ambiguity and inconsistency. When companies give precedence to managing their reputation (by choosing what to and what not to disclose) instead of focusing on what they are doing to limit the spread of misinformation, transparency becomes meaningless.[191] Observers are not able to establish a complete picture of online misinformation trends when the data provided does not correspond to its entirety. Therefore, I propose that standardizing disclosure practices across the industry to improve their efficacy.

Yet, I also recognize that standardization alone is not sufficient. Drawing inspiration from other areas of the law where standardization has proved efficient, I argue that establishing common key-performance indicators is complementary in evaluating platforms' actions and performance over time. Therefore, I suggest introducing specific KPIs to be used by companies to enable assessment of their performance and cross-comparison among different platforms.

Nevertheless, it needs to be stressed that standardization and the adoption of KPIs are only the first steps towards meaningful transparency of decisions aiming at a safer online environment. As I mention throughout this work, many issues still have to be addressed. Firstly, the lack of independent public access and verification of companies' data remains a fundamental obstacle in the fight against misinformation. Without diverting from the trend of

[191] Kevin Roose, *supra* note 180.

only allowing access to selected researchers and restricting the functionalities of APIs, the meaningfulness of transparency of content moderation practices will remain limited.[192] Secondly, finding the right balance between the platform's accountability and freedom to innovate continues to be a challenge for even the most advanced democracies.[193] Thirdly, the tendency to over-rely on technology to solve social and political problems should be countered to avoid being vulnerable to its fallacies.[194]

[192] See Paddy Leerssen, "The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems" (2020) 11:2 European Journal of Law and Technology at 16-18, online: <https://ejlt.org/index.php/ejlt/article/view/786/1012>.
[193] Alexandre De Streel, "Webinar on the Digital Services Act Package: Transparency of content moderation on social media" (22 March 2021), posted on European Audiovisual Observatory, online: YouTube < https://www.youtube.com/watch?v=c0s9nEbEdT0>.
[194] Adam Sinnerich, "Moderation, community, and democracy: Democracy cannot survive algorithmic content moderation", in Tarleton Gillespie et al., *supra* note 68 at 12-14.

6. Bibliography

PRIMARY SOURCES: GOVERNMENT DOCUMENTS

Congressional Research Service, *Free Speech and the Regulation of Social Media Content (R45650)*, by Valery C. Brannon (27 March 2019).

European Union, European Parliament, *Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States*, study (PE 608.864) (LIBE Committee), online: <https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2019)608864>

EU, *Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement*, (Commission SWD 180) (10 September 2020), online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212>.

EU, *Guidance on Strengthening the Code of Practice on Disinformation*, (Commission COM 262) (26 May 2021), online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/redirection/document/76495>.

EU, *Tackling online disinformation: a European Approach*, (Commission COM 236) (26.04.2018), online (pdf): *European Commission* <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&rid=2>.

PRIMARY SOURCES: GERMAN LAW

Netzdurchsetzunggesetz (Network Enforcement Act), (Federal Law Gazette I, p. 3352 ff. Valid as from 1 October 2017).

PRIMARY SOURCES: CANADIAN LAW

Elections Modernization Act, SC 2018, c 31, s 208.1.

PRIMARY SOURCES: FRENCH LAW

Loi organique n° 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1), JO, 23 December 2018.

PRIMARY SOURCES: SINGAPORE LAW

Protection from Online Falsehoods and Manipulation Bill, (Government Gazette, Notification No. B 10, 01 April 2019).

PRIMARY SOURCES: EU LAW

EU, *European Parliament and Council Directive 2013/34/EU of 26 June 2013 on the annual financial statements, consolidated financial statements and related reports of certain types of undertakings, amending Directive 2006/43/EC of the European Parliament and of the Council and repealing Council Directives 78/660/EEC and 83/349/EEC*, [2013] OJ, L 138/19.

European Union, *Code of Practice on Disinformation*, 2018, online (pdf): *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454>

PRIMARY SOURCES: MALAYSIAN LAW

Akta Antiberita Tidak Benar 2018, (Federal Government Gazzette, Adopted on 11 April 2018), online (pdf): https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/106305/130354/F-927153343/MYS106305%20Mys.pdf [Anti-Fake News Act 2018 [Act 803].

PRIMARY SOURCES: US LAW

US, *Breaking the News: Censorship, Suppression, and the 2020 Election*: *Hearing before the Committee on the Judiciary*, (17 November 2020) at 03h:48m ff, online (video): *United States Senate Committee on the Judiciary* <https://www.judiciary.senate.gov/meetings/breaking-the-news-censorship-suppression-and-the-2020-election>

SECONDARY SOURCES: BOOKS

Akriti Gaur, "Towards Policy and Regulatory Approaches for Combating Misinformation in India" in Michael Karanicolas, ed, *Tackling the "Fake" Without Harming the "News": A Paper Series on Regulatory Responses to Misinformation* (Wikimedia/Yale Law School, 2021) 30.

Archon Fung, Mary Graham & David Weil, *Full disclosure: the perils and promise of transparency* (New York: Cambridge Univ. Press, 2007).

Ben Bradford et al, *Report Of The Facebook Data Transparency Advisory Group* (Yale Law School: The Justice Collaboratory, 2019), online (pdf): *Yale Law School* <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf>.

Ben Wagner & Carolina Ferro, *Governance of Digitalization in Europe A contribution to the Exploration Shaping Digital Policy - Towards a Fair Digital Society?*, 1st ed, Barbara Serfozo, ed (Gütersloh, Germany: Bertelsmann Stiftung, 2020).

Ben Wagner, *Global Free Expression - Governing the Boundaries of Internet Content* (Springer International Publishing, 2016).

Daphne Keller & Paddi Leerssen, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily & Joshua A. Tucker, eds, *Social media and democracy: the state of the field, prospects for reform* (Cambridge, United Kingdom; New York, Ny: Cambridge University Press, 2020) 220.

David Kaye, *Speech police: the global struggle to govern the Internet* (New York Columbia Global Reports, 2019).

Don Tapscott & David Ticoll, *Naked corporation: how the age of transparency will revolutionize business* (Toronto: Viking Canada, 2012).

Eleni Kosta & Magdalena Brewczyńska, "Government Access to User Data: Towards More Meaningful Transparency Reports" in Rosa M Ballardini, Petri Kuoppamäki & Olli Pitkänen, eds, *Regulating Industrial Internet Through IPR, Data Protection and Competition Law*

(Kluwer Law International, 2019) , online: SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3601661> .

Garth Jowett & Victoria O'Donnell, *Propaganda & persuasion* (Thousand Oaks: Sage, 2012), online (pdf): <https://hiddenhistorycenter.org/wp-content/uploads/2016/10/PropagandaPersuasion2012.pdf >.

Georgios Terzis et al, *Disinformation and digital media as a challenge for democracy* (Cambridge: Intersentia, 2020).

James Pamment, The EU Code of Practice on Disinformation: Briefing Note for the New EU Commission (Carnegie Endowment for International Peace, 2020).

Marc Liesching et al, *Das NetzDG in der praktischen Anwendung Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes* (Berlin Carl Grossmann Verlag, 2021).

Michael McFaul, ed, *SECURING AMERICAN ELECTIONS Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond.* (Stanford University, 2019), online (pdf): *Stanford University* <http://cs.brown.edu/courses/csci1800/sources/2019_06_06_Stanford_SecuringAmericanElections.pdf>.

Robert Gorwa & Timothy Garton Ash, "Democratic Transparency in the Platform Society" in Nathaniel Persily & Joshua A. Tucker, eds, *Social media and democracy : the state of the field, prospects for reform* (Cambridge, United Kingdom ; New York, Ny: Cambridge University Press, 2020) 286.

Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press, 2019).

Serena Giusti & Elisa Piras, eds, *Democracy and Fake News* (London, UK: Routledge, 2020).

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

Timothy Ash, Robert Gorwa & Danaë Metaxa, *GLASNOST! Nine ways Facebook can make itself a better forum for free speech and democracy* (Reuters Institute, 2019), online (pdf): *Reuters Institute* <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Garton_Ash_et_al_Facebook_report_FINAL_0.pdf>.

Yasha Lange, *Media and Elections* (Council of Europe Publishing, 1999), online (pdf): <https://rm.coe.int/0900001680483b46>.

Wendy W Moe & David A Schweidel, "Why Do We Share Our Opinions?" in *Social Media Intelligence* (Cambridge University Press, 2014) 37.

SECONDARY SOURCES: JOURNAL ARTICLES

Aida Ponce, "The Digital Services Act Package: Reflections on the EU Commission's Policy Options" (2020) ETUI Research Papers, DOI: <10.2139/ssrn.3699389>.

Alan Z Rozenshtein, "Surveillance Intermediaries" (2017) 70 Stanford Law Review 99, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2935321>.

Amélie Heldt, "Reading between the lines and the numbers: an analysis of the first NetzDG reports" (2019) 8:2 Internet Policy Review 1.

Andrea H. Goularta & Ivette K. Muñoza, "Disinformation and post-truth in the context of the Covid-19 pandemic: a study of information practices on Facebook" (2020) 16:2 Liinc em Revista e5395, DOI: <10.18617/liinc.v16i2.5397>.

Andreas Jungherr & Ralph Schroeder, "Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy" (2021) 7:1 Social Media + Society 205630512198892, DOI: <10.1177/2056305121988928>.

Andrej Savin, "The EU Digital Services Act: Towards a More Responsible Internet" (2021) CBS LAW Research Paper No 21-04, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786792>

Anja Bechmann, "Tackling Disinformation and Infodemics Demands Media Policy Changes" (2020) 8:6 Digital Journalism 855.

Archon Fung, "Infotopia: Unleashing the democratic power of transparency" (2013) 41:2 Politics & Society 183, online: <https://journals.sagepub.com/doi/abs/10.1177/0032329213483107>.

Ashley Deeks, "Facebook Unbound?" (2019) 105 Virginia Law Review 1, online: <https://ssrn.com/abstract=3341590>.

Athanasios Andreou et al., "Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations" (Paper delivered at the Network and Distributed System Security Symposium, February 2018), (2018), online (pdf): HAL archives ouvertes <https://hal.archives-ouvertes.fr/hal-01955309/document>.

Ben Wagner et al, "Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act" (2020) FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 261, online: *ACM* <https://dl.acm.org/doi/abs/10.1145/3351095.3372856>.

Benoît Frydman & Isabelle Rorive, "Regulating Internet Content through Intermediaries in Europe and the USA" (2002) 23:1 Zeitschrift für Rechtssoziologie, online: <https://www.degruyter.com/document/doi/10.1515/zfrs-2002-0104/html>.

Bente Kalsnes, "Fake News" (2018) Oxford Research Encyclopedia of Communication, online: *Oxford Research Encyclopedia* <https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-809>.

Blayne Haggarts & Clara Iglesias Keller, "Democratic legitimacy in global platform governance" (Forthcoming) Telecommunications Policy, online: *GIGA Net* <https://www.giga-net.org/2020symposiumPaper/Haggart%20%26%20Keller.pdf?_t=1602675822>.

Briana M. Trifiro & Jennifer Gerson, "Social Media Usage Patterns: Research Note Regarding the Lack of Universal Validated Measures for Active and Passive Use" (2019) 5:2 Social Media + Society, DOI: <10.1177/2056305119848743>.

Cary Coglianese & David Lehr, "Transparency and Algorithmic Governance" (2019) 71 Admin L Rev 1, online: *University of Pennsylvania* <https://scholarship.law.upenn.edu/faculty_scholarship/2123/>.

Chinmayi Arun, "Facebook's Faces" (2021), online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3805210>.

Chris Marsden, Trish Meyer & Ian Brown, “Platform values and democratic elections: How can the law regulate digital disinformation” (2020) 36 Computer Law & Security Review.

Chris Tenove, “Protecting Democracy from Disinformation: Normative Threats and Policy Responses” (2020) 25:3 The International Journal of Press/Politics 517.

Christopher Parsons, “The (In)effectiveness of Voluntarily Produced Transparency Reports” (2017) 58:1 Business & Society 103, online: <https://journals.sagepub.com/doi/full/10.1177/0007650317717957>.

Christos A. Frangonikolopoulos & Ioannis Chapsos, “Explaining the Role and the Impact of the Social Media in the Arab Spring”, (2012) 8:1 GMJ: Mediterranean Edition 10, online (pdf): <https://www.academia.edu/download/30406181/Global_Media_Journal.pdf >

Cynthia Stohl, Michael Stohl & Paul M. Leonardi, “Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age” (2016) 10:5 International Journal of Communication 123, online (pdf): <https://ijoc.org/index.php/ijoc/article/view/4466/1530>.

Daniel Kreiss & Shannon C. Mcgregor, “The ‘Arbiters of What Our Voters See’: Facebook and Google’s Struggle with Policy, Process, and Enforcement around Political Advertising” (2019) 36:4 Political Communication 1, online: <https://www.tandfonline.com/doi/epub/10.1080/10584609.2019.1619639?needAccess=true >.

Daniel Susser, Beate Roessler & Helen F. Nissenbaum, “Online Manipulation: Hidden Influences in a Digital World” (2018) 4:1 Georgetown Law Technology Review, DOI: <10.2139/ssrn.3306006>.

Edda Humprecht, “How Do They Debunk ‘Fake News’? A Cross-National Comparison of Transparency in Fact Checks” (2019) 8:3 Digital Journalism 310.

Emiliana De Blasio & Donatella Selva, “Who Is Responsible for Disinformation? European Approaches to Social Platforms’ Accountability in the Post-Truth Era” (2021) 65:6 American Behavioral Scientist 825, DOI: <10.1177%2F0002764221989784>.

Eugenia Siapera & Paloma Viejo-Otero, “Governing Hate: Facebook and Digital Racism” (2021) 22:2 Television & New Media 112, online: <https://journals.sagepub.com/doi/abs/10.1177/1527476420982232>.

Evelyn Douek, “Facebook’s ‘Oversight Board’: Move Fast with Stable Infrastructure and Humility” (2019) 21 NC JL & TECH 1, online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3365358>.

Evelyn Mary Aswad, "The Future of Freedom of Expression Online" (2018-2019) 17 Duke L & Tech Rev 26.

Fabio Giglietto et al, “It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections” (2020) 23:6 Information, Communication & Society 867, DOI: <10.1080/1369118X.2020.1739732>.

Flavia Durach, Alina Bârgăoanu & Cătălina Nastasiu, “Tackling Disinformation: EU Regulation of the Digital Space” (2020) 20:1 Romanian Journal of European Affairs 5, online: <https://www.ceeol.com/search/article-detail?id=859431>.

Florian Saurwein & Charlotte Spencer-Smith, “Combating Disinformation on Social Media: Multilevel Governance and Distributed Accountability in Europe” (2020) 8:6 Digital Journalism 820, online: <https://doi.org/10.1080/21670811.2020.1765401>.

Frank A. Pasquale, “Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries” (2010) 104:1 Northwestern University Law Review 105, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1686043>.

Giovanni De Gregorio, “The Rise of Digital Constitutionalism in the European Union” (2020) International Journal of Constitutional Law (Forthcoming), online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3506692>.

Halina Szejnwald Brown, Martin de Jong & Teodorina Lessidrenska, “The rise of the Global Reporting Initiative: a case of institutional entrepreneurship” (2009) 18:2 Environmental Politics 182.

Halina Szejnwald Brown, Martin de Jong & Teodorina Lessidrenska, “The rise of the Global Reporting Initiative: a case of institutional entrepreneurship” (2009) 18:2 Environmental Politics 182, DOI: <10.1080/09644010802682551>

Hannah Bloch-Wehba, “Access to Algorithms” (2020) 88:4 Fordham L Rev 1265, online: <https://scholarship.law.tamu.edu/facscholar/1401/>.

Hannah Bloch-Wehba, “Automation in Moderation” (2020) 53 Cornell International Law Journal 41, online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3521619>.

Hannah Bloch-Wehba, “Global Platform Governance: Private Power in the Shadow of the State” (2019) 72:1 SMU L Rev 27, online: *SMU Law Review* <https://scholar.smu.edu/cgi/viewcontent.cgi?article=4778&context=smulr>.

Hans Krause Hansen & Mikkel Flyverbom, “The politics of transparency and the calibration of knowledge in the digital age” (2014) 22:6 Organization 872, online: <https://journals.sagepub.com/doi/full/10.1177/1350508414522315>.

Hunt Allcott & Matthew Gentzkow, “Social Media and Fake News in the 2016 Election” (2017) 31:2 Journal of Economic Perspectives 211, online: <https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.211>.

Hunt Allcott, Matthew Gentzkow & Chuan Yu, “Trends in the diffusion of misinformation on social media” (2019) 6:2 Research & Politics 1, online (pdf): <https://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf>.

Jack M. Balkin, “Free Speech is a Triangle” (2018) 118:7 Colum L Rev 2011, online: *Columbia Law Review* <https://columbialawreview.org/content/free-speech-is-a-triangle/>.

Jakko Kemper & Daan Kolkman, “Transparent to whom? No algorithmic accountability without a critical audience” (2018) 22:14 Information, Communication & Society 2081.

James Grimmelmann, “The Platform is the Message” (2018) 2 Geo L Tech Rev 217, online: *Georgetown Law and Technology Review* <https://georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Grimmelmann-pp-217-33.pdf>.

James Grimmelmann, “The Virtues of Moderation” (2015) 17 Yale JL & Tech 42.

Jan H. Kietzmann et al, “Social media? Get serious! Understanding the functional building blocks of social media” (2011) 54:3 Business Horizons 241–251, online: <https://doi.org/10.1016/j.bushor.2011.01.005>.

Jason Pielemeier, "Disentangling Disinformation: What Makes Regulating Disinformation So Difficult?" (2020) 2020:4 Utah L Rev 917.

Jeffrey Rosen, "The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google" (2013) 6 Constitutional L Rev 35

Jennifer C. Daskal, “Speech Across Borders” (2019) 105:8 Virginia Law Review 1605, online: *Virginia Law Review* <https://www.virginialawreview.org/articles/speech-across-borders/>.

Joachim Haupt, “Facebook futures: Mark Zuckerberg’s discursive construction of a better world” (2021) 23:2 New Media & Society 237, online: <https://journals.sagepub.com/doi/abs/10.1177/1461444820929315>.

Joshua A. Kroll et al, “Accountable Algorithms” (2016) 165 U Pa L Rev 633, online: <https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/>.

Jurgen de Jong & Michiel S. de Vries, “Towards unlimited transparency? Morals and facts concerning leaking to the press by public officials in the Netherlands” (2007) 27:3 Public Administration and Development 215.

Karolina Koc-Michalska et al, “Facebook affordances and citizen engagement during elections: European political parties and their benefit from online strategies?” (2020) Journal of Information Technology & Politics 1, DOI: <10.1080/19331681.2020.1837707>.

Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129:8 Yale LJ 2418.

Kate Klonick, “The New Governors: The People, Rules, and Processes Governing Online Speech” (2017) 131 Harv L Rev 1598, online: <https://ssrn.com/abstract=2937985>.

Katharine Dommett, “Regulating Digital Campaigning: The Need for Precision in Calls for Transparency” (2020) 12:4 Policy & Internet 432.

Kimberly Hall, “Public Penitence: Facebook and the Performance of Apology” (2020) 6:2 Social Media + Society, online: <https://journals.sagepub.com/doi/full/10.1177/2056305120907945>.

Kimberly Hall, “Public Penitence: Facebook and the Performance of Apology” (2020) 6:2 Social Media + Society 205630512090794, online: <https://journals.sagepub.com/doi/full/10.1177/2056305120907945>.

Kris Hartley & Minh Khuong Vu, “Fighting fake news in the COVID-19 era: policy insights from an equilibrium model” (2020) 53 Policy Sci, online: <https://doi.org/10.1007/s11077-020-09405-z>.

Kristen E. Eichensehr, “Digital Switzerlands” (2018) 167 U PA L Rev 665, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3205368>.

Kyle Langvardt, “Regulating Online Content Moderation” (2017) 106:5 Georgetown Law Journal 1353, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3024739>.

Lavinia Marin, “Sharing (mis)information on social networking sites. An exploration of the norms for distributing content authored by others” (2021) Ethics and Information Technology, DOI: 0.1007/s10676-021-09578-y

Lindsay Stirton & Martin Lodge, “Transparency Mechanisms: Building Publicness into Public Services” (2001) 28:4 Journal of Law and Society 471.

Manuel Puppis, “Media Governance: A New Concept for the Analysis of Media Policy and Regulation” (2010) 3:2 Communication, Culture & Critique 134, online: <https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1753-9137.2010.01063.x>.

Marco A. Ruedriger & Amaro Grassi, “Desinformação on-line e processos políticos: a circulação de links sobre desconfiança no sistema eleitoral brasileiro no Facebook e no

YouTube (2014-2020)" (2020) FGV Policy Paper, online (pdf): <https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/30085/%5bPT%5d%20Estudo%201%20%281%29.pdf?sequence=1&isAllowed=y>.

Maria Cucciniello, Gregory A. Porumbescu & Stephan Grimmelikhuijsen, "25 Years of Transparency Research: Evidence and Future Directions" (2016) 77:1 Public Administration Review 32, online: *Wiley Online Library* <https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.12685>.

Matthew P. Hooker, "Censorship, Free Speech & Facebook: Applying the First Amendment to Social Media Platforms via the Public Function Exception" (2019) 15:1 Washington J of L, Technology & Arts 36.

Michael Karanicolas, "A FOIA for Facebook: Meaningful Transparency for Online Platforms" (Paper delivered at the FESC of the Floyd Abrams Institute for Freedom of Expression, Yale Law School, 30 April 2021) [unpublished].

Michael Karanicolas, "Even in a Pandemic, Sunlight is the Best Disinfectant: COVID-19 and Global Freedom of Expression" (2020) 20 Or Rev Int'l L 101, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3726892>.

Mike Ananny & Kate Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" (2016) 20:3 New Media & Soc'y 973, online: *Sage Journals* <https://journals.sagepub.com/doi/full/10.1177/1461444816676645>.

Mikkel Flyverbom, "Digital Age Transparency: Mediation and the Management of Visibilities" (2016) 10:0 International Journal of Communication 13, online: <https://ijoc.org/index.php/ijoc/article/view/4490>.

Mikkel Flyverbom, "Sunlight in cyberspace? On transparency as a form of ordering" (2015) 18:2 European Journal of Social Theory 168, online: <https://journals.sagepub.com/doi/10.1177/1368431014555258>.

Mikkel Flyverbom, Lars Thøger Christensen & Hans Krause Hansen, "The Transparency–Power Nexus" (2015) 29:3 Management Communication Quarterly 385, online: <https://journals.sagepub.com/doi/10.1177/0893318915593116 >

Min Jiang & King-Wa Fu, "Chinese Social Media and Big Data: Big Data, Big Brother, Big Profit?" (2018) 10:4 Policy & Internet 372.

Molly K. Land, "Against Privatized Censorship: Proposals for Responsible Delegation" (Forthcoming) 60:2 Virginia Journal of International Law 363, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3442184>

Mônica Chaves & Adriana Braga, "The agenda of disinformation: 'fake news' and membership categorization analysis in the 2018 Brazilian presidential elections" (2019) 15:3 Brazilian Journalism Research 474, DOI: <10.25200/BJR.v15n3.2019.1187>.

Monika Bickert, "Charting a Way Forward: Online Content Regulation" (2020), online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf >

Monika Zalnieriute, "'TransparencyWashing'in the Digital Age: A Corporate Agenda of Procedural Fetishism" (2021) 8:1 Critical Analysis of Law 140.

Monika Zalnieriute, "'TransparencyWashing'in the Digital Age: A Corporate Agenda of Procedural Fetishism" (2021) 8:1 Critical Analysis of Law 140, online (pdf): *University of Toronto* https://cal.library.utoronto.ca/index.php/cal/article/view/36284/27587

Nicolas P. Suzor et al, “What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation” (2019) 13 Int’l J of Comm 1526, online: <https://ijoc.org/index.php/ijoc/article/view/9736>.

Oana Brindusa Albu & Mikkel Flyverbom, “Organizational Transparency: Conceptualizations, Conditions, and Consequences” (2016) 58:2 Business & Society 268–297, online: <https://journals.sagepub.com/doi/10.1177/0007650316659851>.

Ondina Gabrovec Mei, “CSR and Social Reporting: Moving Towards Standardization” (2013) [unpublished], online (pdf): *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2239658&download=yes

Pablo Ornelas Rosa, Aknaton Toczek Souza & Giovane M Camargo, “Perspectividade política e produção de desinformação nas eleições brasileiras de 2018” (2021) 8:3 Agenda Política 163, DOI: <10.31990/agenda.2020.3.6>

Paddy Leerssen, “The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems” (2020) 11:2 European Journal of Law and Technology, online: <https://ejlt.org/index.php/ejlt/article/view/786/1012>.

Panom Gunawong, “Open Government and Social Media” (2014) 33:5 Social Science Computer Review 587.

Patrícia Rossini et al., “Explaining Dysfunctional Information Sharing On Whatsapp And Facebook In Brazil” (paper delivered at the 21st Annual Conference of the Association of Internet Researchers, 27-31 October 2020), [unpublished].

Paul Butcher, “Disinformation and democracy: The home front in the information war” (discussion paper for the European Politics And Institutions Programme, 30 January 2019) European Policy Center, online (pdf): < https://www.epc.eu/content/PDF/2019/190130_Disinformationdemocracy_PB.pdf>.

Paul-Jasper Dittrich, “Tackling the spread of disinformation: Why a co-regulatory approach is the right way forward for the EU” (2019) Bertelsmann Stiftung, online: <http://aei.pitt.edu/102500/1/2019.dec.pdf>.

Petros Iosifidis & Nicholas Nicoli, “The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation” (2019) 82:1 International Communication Gazette 60.

Philippe Verduyn et al, “Do Social Network Sites Enhance or Undermine Subjective Well-Being? A Critical Review” (2017) 11:1 Social Issues and Policy Review 274, DOI: <10.1111/sipr.12033>

Rachael Craufurd Smith, “Fake news, French Law and democratic legitimacy: lessons for the United Kingdom?” (2019) 11:1 Journal of Media Law 1, DOI: <10.1080/17577632.2019.1679424>.

Robert Faris et al., “Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election” (2017) Berkman Klein Center Research Publication 2017-6, online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3019414>.

Robert Gorwa, “The platform governance triangle: conceptualising the informal regulation of online content” (2019) 8:2 Internet Policy Review.

Robert Gorwa, “What is platform governance?” (2019) 22:6 Information, Communication & Society 854.

Robert Gorwa, Reuben Binns & Christian Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance” (2020) 7:1 Big Data & Society, online: *Sage Journals* <https://journals.sagepub.com/doi/full/10.1177/2053951719897945>.

Samantha Bradshaw, The Social Media Challenge for Democracy: Propaganda and Disinformation in a Platform Society, (PhD Thesis, University of Oxford, 2020) [unpublished], online: *Oxford University Research Archive* <https://ora.ox.ac.uk/objects/uuid:e75e4796-d614-454b-b2e2-df6b8659e610>.

Sarah Myers West, “Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms” (2018) 20:11 New Media & Society 4366, online: <https://journals.sagepub.com/doi/full/10.1177/1461444818773059>.

Soroush Vosoughi, Deb Roy & Sinan Aral, “The spread of true and false news online” (2018) 359:6380 Science 1146–1151, DOI: 10.1126/science.aap9559

Tarleton Gillespie et al, “Expanding the debate about content moderation: scholarly research agenda in the coming policy debates” (2020) 9:4 Internet Policy Review.

Tatiana M. S. G. Dourado, *Fake news na eleição presidencial de 2018 no Brasil*, (PhD Thesis, Universidade Federal da Bahia, 2020) [unpublished].

Tim Hwang, “Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment” (2017) SSRN Electronic Journal, online: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442> [unpublished]

Timothy Caulfield, “Does Debunking Work? Correcting COVID-19 Misinformation on Social Media” in Colleen M. Flood et al, eds., *Vulnerable: The Law, Policy and Ethics of COVID-19* (Univeristy of Ottawa Press, 2020).

Tobias R Keller et al, *#ARSONEMERGENCY: Climate Change Disinformation During the Australian Bushfire Season 2019-2020* (paper delivered at the 21st Annual Conference of the Association of Internet Researchers, 27-31 October 2020), [unpublished].

SECONDARY SOURCES: REPORTS

Australian Communications and Media Authority, *Misinformation and news quality on digital platforms in Australia* (June 2020), online (pdf): *ACMA* <https://www.acma.gov.au/sites/default/files/2020-06/Misinformation%20and%20news%20quality%20position%20paper.pdf>.

Avaaz, *Facebook: From Election to Insurrection - How Facebook Failed Voters and Nearly Set Democracy Aflame* (18 March 2021), online: *Avaaz* https://secure.avaaz.org/campaign/en/facebook_election_insurrection/

Avaaz, *Far Right Networks of Deception* (22/05/2019), online (pdf): *Avaaz* <https://avaazimages.avaaz.org/Avaaz%20Report%20Network%20Deception%2020190522.pdf>

Center for an Informed Public et al., *The Long Fuse: Misinformation and the 2020 Election, 2021*, (2021), online (pdf): *Stanford Digital Repository* <https://purl.stanford.edu/tr171zs0069>.

Data & Society, *Dead Reckoning. Navigating Content Moderation After "Fake News"*, by Robyn Caplan, Lauren Hanson & Joan Donovan (February 2018), online (pdf): *Data & Society* <https://datasociety.net/pubs/oh/DataAndSociety_Dead_Reckoning_2018.pdf>.

Data & Society, *Media manipulation and disinformation online*, by Alice Marwick & Rebecca Lewis (2017), online (pdf): *Data & Society* <https://datasociety.net/wp-content/uploads/2017/05/DataAndSociety_MediaManipulationAndDisinformationOnline-1.pdf>

European Commission, *Study for the Assessment of the Implementation of the Code of Practice on Disinformation*, by Iva Plasilova et al. (May 2020), online: *European Commission* <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=66649>

European Parliament, *Study on The impact of disinformation on democratic processes and human rights in the world*, by Carme Colomina, Héctor Sánchez Margalef & Richard Youngs (April 2021), online (pdf): *European Parliament* < https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf>.

European Commission, *A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation*, by Madeleine De Cock Buning (March 2018), online (pdf): *European University Institute* <http://diana-n.iue.it:8080/handle/1814/70297>

Facebook Inc., *A Look at Facebook and the US 2020 Elections* (December 2020), online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2020/12/US-2020-Elections-Report.pdf>.

Facebook Inc., *April 2021 Coordinated Inauthentic Behavior Report* (May 2021), online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2021/05/April-2021-CIB-Report.pdf>

Facebook Inc., *Facebook Baseline Report on Implementation of the Code of Practice on Disinformation* (May 2019), online (pdf): *European Commission* <https://ec.europa.eu/information_society/newsroom/image/document/2019-5/facebook_baseline_report_on_implementation_of_the_code_of_practice_on_disinformation_CF161D11-9A54-3E27-65D58168CAC40050_56991.pdf>

Facebook Inc., *Facebook reports on implementation of the Code of Practice on Disinformation – May report*, (22 May 2019), online (pdf): *European Commission* https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60041

Facebook Inc., *Facebook Transparency Report*s (2021), online: *Facebook* <https://transparency.facebook.com/>.

Facebook Inc., *Information Operations and Facebook*, by Jen Weedon, William Nuland & Alex Stamos (27 April 2017), online (pdf): *Facebook* <https://about.fb.com/wp-content/uploads/2017/04/facebook-and-information-operations-v1.pdf>

Facebook Inc., *June 2021 Coordinated Inauthentic Behavior Report (July 2021)*, online (pdf): <https://about.fb.com/wp-content/uploads/2021/07/June-2021-CIB-Report-Final.pdf>.

Freedom House, *Freedom on the Net 2017: Manipulating Social Media to Undermine Democracy*, by Sanja Kelly et al. (February 2017), online: *Freedom House* < https://freedomhouse.org/report/freedom-net/2017/manipulating-social-media-undermine-democracy>.

Google LLC, “Google Transparency Report - Government requests to remove content” (2021), online: *Google* <https://transparencyreport.google.com/government-removals/overview?hl=en>.

Google LLC., *Google Transparency Report* (2020), online: *Google* <https://transparencyreport.google.com/youtube-policy/removals>.

Justice Collaboratory, *Report Of The Facebook Data Transparency Advisory Group*, by Ben Bradford et al (April 2019), online (pdf): *Yale Law School* <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf >

KPMG, *The KPMG Survey of Corporate Responsibility Reporting 2017*, by Jose Luis Blasco & Adrian King (2017), online (pdf): *KPMG* https://assets.kpmg/content/dam/kpmg/xx/pdf/2017/10/kpmg-survey-of-corporate-responsibility-reporting-2017.pdf

Online Civic Culture Centre, *News Sharing On UK Social Media: Misinformation, Disinformation, And Correction*, by Andrew Chadwick & Cristian Vaccari (2019), online: *Loughborough University* < https://www.lboro.ac.uk/research/online-civic-culture-centre/news-events/articles/o3c-1-survey-report-news-sharing-misinformation/>.

Oxford Internet Institute, *The Market of Disifinformation*, by Stacie Hoffmann, Emily Taylor & Samantha Bradshaw (October 2019), online (pdf): *Oxford Internet Institute* <https://oxil.uk/publications/oxtec-market-of-disinformation/OxTEC-The-Market-of-Disinformation.pdf>

Stanford University, Report of the Working Group on Platform Scale, by Francis Fukuyama et al. (2020), online (pdf): *Stanford University* <https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/platform_scale_whitepaper_-cpc-pacs.pdf>.

Twitter Inc., *Information Operations Report* (2021), online: *Twitter* <https://transparency.twitter.com/en/reports/information-operations.html>.

Twitter Inc., *Platform Manipulation Report* (11 January 2021), online: *Twitter* <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun>.

Twitter Inc., *Rules Enforcement Report* (11 January 2021), online: *Twitter* <https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun>.

SECONDARY SOURCES: WEBSITES

“Propaganda Eleitoral na Internet” (2018) at 11, online (pdf): *Justiça Eleitoral* https://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet/rybena_pdf?file=https://www.justicaeleitoral.jus.br/arquivos/propaganda-eleitoral-na-internet/at_download/file [“Electoral Advertisement on the Internet”, translated by the author](

Adam Tornes & Leanne Trujillo, “Enabling the future of academic research with the Twitter API”, (26 January 2021), online: *Twitter* <https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api>.

Aleksandra Kuczerawy, “The Code of Conduct on Online Hate Speech: an example of state interference by proxy? - CITIP blog” (20 July 2016), online: *CITIP blog*

<https://www.law.kuleuven.be/citip/blog/the-code-of-conduct-on-online-hate-speech-an-example-of-state-interference-by-proxy/>.

Amy Watson, "Europe: trust in the written press by country 2021 | Statista", (21 May 2021), online: *Statista* <https://www.statista.com/statistics/454403/europe-trust-in-the-written-press-by-country/>.

Anika Geisel, "Protecting the European Parliament Elections - About Facebook" (28 January 2019), online: *Facebook* <https://about.fb.com/news/2019/01/european-parliament-elections/>.

Anika Geisel, "Protecting the European Parliament Elections" (28 January 2019), online: *Facebook* <https://about.fb.com/news/2019/01/european-parliament-elections/>.

Anti-Defamation League et al., "Facebook's Suspension of Donald J. Trump" (12 February 2021), online (pdf): *Free Press* <https://www.freepress.net/sites/default/files/2021-02/comment_on_trump_suspension_to_oversight_board.pdf>

Antoine Grégoire, "CIB Detection Tree: 2nd Branch | EU DisinfoLab", (14 June 2021), online: *EU DisinfoLab* <https://www.disinfo.eu/publications/cib-detection-tree2/>.

Asher Schechter, "Brazil's Election Is Yet Another Indication That Facebook Is Too Big to Manage - ProMarket" (31 October 2018), online: *ProMarket - University of Chicago Booth* <https://promarket.org/2018/10/31/brazils-election-is-yet-another-indication-that-facebook-is-too-big-to-manage/>.

Ben Wagner et al, "Auditing Big Tech: Tackling Disinformation and the EU Digital Services Act" (2021), online (pdf): *Enabling Digital Rights and Governance* <https://enabling-digital.eu/wp-content/uploads/2021/02/Auditing_big_tech_Final.pdf>.

Center for Democracy and Technology, "Comments to Facebook Oversight Board" (11 February 2021), online (pdf): *Center for Democracy and Technology* <https://cdt.org/wp-content/uploads/2021/02/CDT-comments-to-FB-Oversight-Board-on-2021-001-FB-FBR.pdf>

Craig Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook" (16 November 2016), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>

Craig Silverman, Ryan Mac & Pranav Dixit, "Whistleblower Says Facebook Ignored Global Political Manipulation" (14 September 2020), online: *BuzzFeed News* <https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo>.

Daniel Funke & Daniela Flamini, "A guide to anti-misinformation actions around the world - Poynter" (14 August 2019), online: *Poynter* <https://www.poynter.org/ifcn/anti-misinformation-actions/>.

David Kaye, "The Republic of Facebook" (6 May 2020), online: *Just Security* <https://www.justsecurity.org/70035/the-republic-of-facebook/>.

Ed Bracho-Polanco, "How Jair Bolsonaro used 'fake news' to win power" (8 January 2019), online: *The Conversation* <https://theconversation.com/how-jair-bolsonaro-used-fake-news-to-win-power-109343>.

Elisa Shearer, “More than eight-in-ten Americans get news from digital devices” (12 January 2021), online: *Pew Research Center* <https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/>.

European Commission, “Code of Practice on Disinformation - Intermediate Targeted Monitoring - May Reports” (2019), online: *European Commission* <https://digital-strategy.ec.europa.eu/en/news/last-intermediate-results-eu-code-practice-against-disinformation>.

Facebook Inc., “Combatendo a desinformação para proteger a eleição no Brasil - Sobre o Facebook” (23 October 2018), online: *Facebook* <https://about.fb.com/br/news/2018/10/combatendo-a-desinformacao-para-proteger-a-eleicao-no-brasil/>.

Facebook Inc., “Facebook Community Standards” (2021), online: *Facebook* <https://www.facebook.com/communitystandards/>.

Felipe Pontes, “TSE assina memorando com Facebook e Google contra fake news”, (28 June 2018), online: *Agência Brasil* <https://agenciabrasil.ebc.com.br/justica/noticia/2018-06/tse-assina-memorando-com-facebook-e-google-contra-fake-news>. [“Supreme Electoral Tribunal signs a memorandum of understanding with Facebook and Google against fake-news”, translated by the author]

Felix Salmon, “Media trust hits new low”, (21 January 2021), online: *Axios* <https://www.axios.com/media-trust-crisis-2bf0ec1c-00c0-4901-9069-e26b21c283a9.html>

George Serafeim, “Social-Impact Efforts That Create Real Value” (September 2020), online: *Harvard Business Review* <https://hbr.org/2020/09/social-impact-efforts-that-create-real-value>.

Google LLC., “YouTube Data API Overview” (2012), online: *Google Developers* <https://developers.google.com/youtube/v3/getting-started?hl=it>.

Google LLC., *YouTube Community Guidelines & Policies* (2021), online: *YouTube* <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/>.

Google LLC., “Spam, deceptive practices, & scams policies - YouTube Help” (2021),online: *Google* <https://support.google.com/youtube/answer/2801973?hl=en>

Guy Rosen & Monika Bickert, “Our Response to the Violence in Washington - About Facebook”, (7 January 2021), online: *Facebook* <https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/>.

Guy Rosen et al, “Helping to Protect the 2020 US Elections - About Facebook” (21 October 2019), online: *Facebook* <https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/>.

Guy Rosen, “An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19” (16 April 2020), online: *Facebook* <https://about.fb.com/news/2020/04/covid-19-misinfo-update/>.

Guy Rosen, “Preparing for Election Day”, (7 October 2020), online: *Facebook* <https://about.fb.com/news/2020/10/preparing-for-election-day/>.

Guy Rosen, “Protecting Facebook Live From Abuse and Investing in Manipulated Media Research - About Facebook” (15 May 2019), online: *Facebook* <https://about.fb.com/news/2019/05/protecting-live-from-abuse/>.

Irene Pasquetto & Briony Swire-Thompson, "Tackling misinformation: What researchers could do with social media data", (9 December 2020), online: *Harvard Kennedy School Misinformation Review* <https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/>.

Jenna Hand, "'Fake news' laws, privacy & free speech on trial: Government overreach in the infodemic?" (12 August 2020), online: *First Draft* <https://firstdraftnews.org/latest/fake-news-laws-privacy-free-speech-on-trial-government-overreach-in-the-infodemic/>.

Katie Harbath, "Protegendo as eleições no Brasil" (24 July 2018), online: *Facebook* <https://about.fb.com/br/news/2018/07/protegendo-as-eleicoes-no-brasil/>.

Kleis Nielsen, Robert Gorwa & Madeleine De Cock Buning, "What can be done? Digital Media Policy Options for Europe (and beyond)" (25 November 2019), online: *Reuters Institute for the Study of Journalism* <https://reutersinstitute.politics.ox.ac.uk/what-can-be-done-digital-media-policy-options-europe-and-beyond>.

Laura Elderson et al, "Far-right news sources on Facebook more engaging" (3 March 2021), online (Medium): *Cybersecurity for Democracy* <https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90>.

Maria A Golino, "Algorithms in Social Media Platforms" (24 April 2021), online: *Institute for Internet & Just Society* <https://www.internetjustsociety.org/algorithms-in-social-media-platforms>.

Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process - About Facebook", (24 April 2018), online: *Facebook* <https://about.fb.com/news/2018/04/comprehensive-community-standards/>.

Oana Goga, "Facebook's 'transparency' efforts hide key reasons for showing ads" (15 May 2019), online: *The Conversation* <https://theconversation.com/facebooks-transparency-efforts-hide-key-reasons-for-showing-ads-115790>.

Rim-Sarah Alouane, "Macron's Fake News Solution Is a Problem" (29 May 2018), online: *Foreign Policy* <https://foreignpolicy.com/2018/05/29/macrons-fake-news-solution-is-a-problem/>.

Sara Fischer, "'Unreliable' news sources got more traction in 2020" (22 December 2020), online: *Axios* <https://www.axios.com/unreliable-news-sources-social-media-engagement-297bf046-c1b0-4e69-9875-05443b1dca73.html>.

Talal Rafi, "Council Post: Why Corporate Strategies Should Be Focused On Sustainability", (10 February 2021), online: *Forbes* <https://www.forbes.com/sites/forbesbusinesscouncil/2021/02/10/why-corporate-strategies-should-be-focused-on-sustainability/?sh=1d46193a7e9f>.

Tarleton Gillespie, "The Platform Metaphor, Revisited", (24 August 2017), online: *HIIG* <https://www.hiig.de/en/the-platform-metaphor-revisited/>.

TikTok, "Application Guidelines", (2021), online: *TikTok for Developers* <https://developers.tiktok.com/doc/getting-started-faq>.

TikTok, "TikTok Community Guidelines", (December 2020), online: *TikTok* <https://www.tiktok.com/community-guidelines?lang=en#37>.

Twitter Inc., "Apply for access – Twitter Developers", (2021), online: *Twitter* <https://developer.twitter.com/en/apply-for-access>.

Twitter Inc., “Permanent suspension of @realDonaldTrump”, (8 January 2021), online: *Twitter* <https://blog.twitter.com/en_us/topics/company/2020/suspension.html>.

Twitter Inc., “Platform manipulation and spam policy”, (September 2020), online: *Twitter* <https://help.twitter.com/en/rules-and-policies/platform-manipulation>.

Twitter Inc., “Twitter Rules and Policies”, (2021), online: *Twitter* <https://help.twitter.com/en/rules-and-policies>.

Twitter Inc., “Twitter Transparency Center”, online: *Twitter* <https://transparency.twitter.com/>.

Twitter Safety Team, “Disclosing networks of state-linked information operations”, (23 February 2021), online: *Twitter* <https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-.html>.

Vicki Jackson & Martha Minow, “Facebook Suspended Trump. The Oversight Board Shouldn’t Let Him Back.”, (8 March 2021), online: *Lawfare* <https://www.lawfareblog.com/facebook-suspended-trump-oversight-board-shouldnt-let-him-back>.

SECONDARY SOURCES: NEWSPAPERS

Angelique Chrisafis, “French MPs criticise ‘hasty and ineffective’ fake news law”, *the Guardian* (8 June 2018), online: <https://www.theguardian.com/world/2018/jun/07/france-macron-fake-news-law-criticised-parliament>.

Daniel Boffey, “EU disputes Facebook’s claims of progress against fake accounts”, *The Guardian* (29 October 2019), online: <https://www.theguardian.com/world/2019/oct/29/europe-accuses-facebook-of-being-slow-to-remove-fake-accounts>.

David Shepardson, “U.S. panel asks FBI to review role of Parler in Jan. 6 Capitol attack”, (21 January 2021), *Reuters* online: <https://www.reuters.com/article/us-usa-trump-parler-idUSKBN29Q2FS>.

Elizabeth Culliford & Jeffrey Dastin, “Parler CEO says social media app, favored by Trump supporters, may not return”, *Reuters* (13 January 2021), online: <https://www.reuters.com/technology/exclusive-parler-ceo-says-social-media-app-favored-by-trump-supporters-may-not-2021-01-13/>.

Elle Hunt, “What is fake news? How to spot it and what you can do to stop it”, *The Guardian* (17 December 2016), online: <https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>.

Emily Bazelon, “Why Is Big Tech Policing Speech? Because the Government Isn’t”, *The New York Times* (26 January 2021), online: <https://www.nytimes.com/2021/01/26/magazine/free-speech-tech.html>.

Hayley Tsukayama, “Facebook releases first report on world governments’ data requests”, *The Washington Post* (27 August 2013), online: <https://www.washingtonpost.com/business/technology/facebook-releases-first-report-on-world-governments-data-requests/2013/08/27/40e2d396-0f24-11e3-8cdd-bcdc09410972_story.html>.

Jeff Horwitz & Deepa Seetharaman, “Facebook Executives Shut Down Efforts to Make the Site Less Divisive”, *The Wall Street Journal* (26 May 2020), online: <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>.

Kevin Roose, “Inside Facebook’s Data Wars”, The New York Times (14 July 2021), online: <https://www.nytimes.com/2021/07/14/technology/facebook-data.html>.

Linda Kinstler, “The Atlantic”, (18 May 2018), online: *Germany’s Attempt to Fix Facebook* <https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/>.

Meghan Mandavia, “Social media to join hands to fight fake news, hate speech”, *The Economic Times* (19 February 2020), online: <https://economictimes.indiatimes.com/tech/internet/social-media-to-join-hands-to-fight-fake-news-hate-speech/articleshow/74200542.cms?from=mdr>.

Megan A Brown et al, “Twitter put warning labels on hundreds of thousands of tweets. Our research examined which worked best.”, *the Washington Post* (9 December 2020), online: <https://www.washingtonpost.com/politics/2020/12/09/twitter-put-warning-labels-hundreds-thousands-tweets-our-research-examined-which-worked-best/>.

Mike Isaac & Kellen Browning, “Fact-Checked on Facebook and Twitter, Conservatives Switch Their Apps”, *The New York Times* (11 November 2021), online: <https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html>.

Newley Purnell & Jeff Horwitz, “Facebook’s Hate-Speech Rules Collide With Indian Politics”, *Wall Street Journal* (14 August 2020), online: <https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346>.

Nicholas Confessore, “Cambridge Analytica and Facebook: The Scandal and the Fallout So Far (Published 2018)”, *The New York Times* (4 April 2021), online: <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>.

Peter Cunliffe-Jones, “Europe’s latest export: A bad disinformation strategy”, *Politico* (7 June 2021), online: <https://www.politico.eu/article/europe-bad-disinformation-strategy-digital-services-act-dsa/amp/>.

Sheera Frenkel & Mike Isaac, “Inside Facebook’s Election ‘War Room’”, *The New York Times* (19 September 2018), online: <https://www.nytimes.com/2018/09/19/technology/facebook-election-war-room.html>.

Steven Erlanger, “‘Fake News,’ Trump’s Obsession, Is Now a Cudgel for Strongmen (Published 2017)”, *The New York Times* (12 December 2017), online: <https://www.nytimes.com/2017/12/12/world/europe/trump-fake-news-dictators.html>.

Vera Bergengruen & Billy Perrigo, “Facebook Acted Too Late to Tackle Misinformation on 2020 Election, Report Finds”, *Time* (23 March 2021), online: <https://time.com/5949210/facebook-misinformation-2020-election-report/>.

Abby Ohlheiser, “Facebook backs down, will no longer censor the iconic ‘Napalm Girl’ war photo” *The Washington Post* (9 September 2016), online: <https://www.washingtonpost.com/news/the-intersect/wp/2016/09/09/abusing-your-power-mark-zuckerberg-slammed-after-facebook-censors-vietnam-war-photo/>.

Lucas Vidigal, Gabriela Sarmento & Cida Alves, "Candidatos destinam 1,6% dos gastos da eleição de 2018 para anúncio online, aponta balanço parcial", *G1* (18 September 2018), online: <https://g1.globo.com/politica/eleicoes/2018/noticia/2018/09/18/candidatos-destinam-16-dos-gastos-da-eleicao-de-2018-para-anuncio-online-aponta-balanco-parcial.ghtml>. [translated by the author]

Deepa Seetharaman & Jeff Horwitz, "Facebook Touted Its Progress in Brazil Elections. Internally There Were Doubts.", *The Wall Street Journal* (30 August 2019), online: <https://www.wsj.com/articles/facebook-said-it-aced-brazil-elections-internally-there-were-doubts-11567157403>.

Alexis C Madrigal, "What Facebook Did to American Democracy", *The Atlantic* (12 October 2017), online: <https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/>.

Olivia Solon, "Facebook's failure: did fake news and polarized politics get Trump elected?" *The Guardian* (10 November 2016), online: <https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories>.

Tony Romm & Kurt Wagner, "Facebook admits 'malicious actors' spread misinformation during the 2016 U.S. election", *Vox* (28 April 2017), online: <https://www.vox.com/2017/4/28/15476142/facebook-report-trump-clinton-russia-us-presidential-election>.

Sheera Frenkel & Katie Benner, "To Stir Discord in 2016, Russians Turned Most Often to Facebook", *The New York Times* (18 February 2018), online: <https://www.nytimes.com/2018/02/17/technology/indictment-russian-tech-facebook.html>.

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

Issie Lapowsky, "Here's How Facebook Actually Won Trump the Presidency", *Wired* (15 November 2016), online: <https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/>.

Rachel Lerman, "Facebook says it labeled 180 million debunked posts ahead of the election", *Washington Post* (19 November 2020), online: <https://www.washingtonpost.com/technology/2020/11/19/facebook-election-warning-labels/>.

"Key Performance Indicators Listed by Sector" (2016), online (pdf): *IntraFocus* <https://static.intrafocus.com/uploads/2016/02/Key-Performance-Indicators-by-Sector.pdf>

"Merkel kritisiert Twitter-Sperre für Trump",*Tagesspiegel* (11 January 2021), online: <https://www.tagesspiegel.de/politik/meinungsfreiheit-von-elementarer-bedeutung-merkel-kritisiert-twitter-sperre-fuer-trump/26786886.html>.

SECONDARY SOURCES: WORKING PAPERS

Craig Matasick, Carlotta Alfonsi & Antonio Bellantoni, "Governance responses to disinformation: How open government principles can inform policy options" (2020) OECD Working Papers on Public Governance No. 39, DOI: <10.1787/d6237c85-en>

Federica Liberini et al., " Politics in the Facebook Era - Evidence from the 2016 US Presidential Elections" (2020) CESifo Working Paper No. 8235, online:*SSRN* <https://ssrn.com/abstract=3584086>.

Garrett Morrow et al, "The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation" (2020) Northeastern University Ethics Institute Working Paper, online: *SSRN* <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3742120>

Maroussia Lévesque, "Applying the Un Guiding Principles on Business and Human Rights to Online Content Moderation" (19 February 2021), Working Paper, DOI: <10.2139/ssrn.3789311>

Matthias C. Kettemann & Wolfgang Schulz, "Setting Rules for 2.7 Billion: a (First) Look into Facebook's Norm-Making System; Results of a Pilot Study" (2020) Working Papers of the Hans-Bredow-Institut, online (pdf): <https://www.ssoar.info/ssoar/bitstream/handle/document/71724/ssoar-2020-kettemann_et_al-Setting_Rules_for_27_Billion.pdf?sequence=4&isAllowed=y&lnkname=ssoar-2020-kettemann_et_al-Setting_Rules_for_27_Billion.pdf>.

Philip N. Howard et al., "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" (2011), PITPI Working Paper 2011.1, DOI: <10.2139/ssrn.2595096>.

Sergei Hovyadinov, "Toward a More Meaningful Transparency: Examining Twitter, Google, and Facebook's Transparency Reporting and Removal Practices in Russia" (2020) Working Paper.

SECONDARY SOURCES: SOCIAL MEDIA

Alexej Navalny, "1. I think that the ban of Donald Trump on Twitter is an unacceptable act of censorship (THREAD)" (9 January 2021), online: *Twitter* https://twitter.com/navalny/status/1347969772177264644?s=20

Alexandre De Streel, "Webinar on the Digital Services Act Package: Transparency of content moderation on social media" (22 March 2021), posted on European Audiovisual Observatory, online: YouTube < https://www.youtube.com/watch?v=c0s9nEbEdT0>.

Mark Zuckerberg, "The US elections are just two months away …" (3 September 2020), posted on *Mark Zuckerberg*, online: *Facebook* <https://www.facebook.com/zuck/posts/10112270823363411>.

Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement" (5 May 2021), posted on *Mark Zuckerberg*, online: *Facebook* https://www.facebook.com/notes/751449002072082/.

Pranav Dixit, "Crucially, however, Twitter refused to respond to questions I asked …" (21 May 2021 06:16), online: *Twitter* <https://mobile.twitter.com/PranavDixit/status/1395730277981294598>.