Combating Disinformation

Platform (un)accountability. Reviewing Platform Responses to the Global Disinfodemic One Year Onward

Trisha Meyer, Alexandre Alaphilippe

Abstract: This chapter compares Facebook, Google, TikTok and Twitter's responses to COVID-19 and US elections-related disinformation in 2020, furthering our understanding of often opaque moderation practices. Most prominently, online platforms heavily emphasized amplification of credible information, including through provision of free advertising space. They also rapidly and regularly expanded their policies in order to ban, remove, demote or label disinformation as harmful but not illegal. In 2020, the editorial role of online platforms became visible as never before. Their ability to react quickly is both encouraging and worrying, if not accompanied by a known hierarchy of principles and stringent transparency and review measures.

Keywords: content moderation; platform power; COVID-19; US 2020 Elections; disinformation; Facebook; Google; TikTok; Twitter

1. Introduction

One year ago, the COVID-19 virus brought economies and societies to a screeching halt. A global health pandemic ensued. One year later, as vaccinations roll out, we hope for return to a 'new normal', with a renewed appreciation of the need for social connection in our lives. In our isolation, community has proven more important than ever.[1]

Parallel to the spread of the virus has been the spread of disinformation, which Posetti, Bontcheva et.al. describe as a 'disinfodemic' in their ITU/UNESCO study on balancing responses to disinformation with freedom

1 Jonathan Sacks, *Morality: Restoring the Common Good in Divided Times* (New York: Basic Books, 2020).

of expression, media and information literacy, and critical independent journalism.[2]

During this health and information pandemic, online platforms are under intense scrutiny to tackle the disinfodemic rampant on their services. Their terms of service, community guidelines, as well as national legislation, seek to dissuade users from posting illegal content – and increasingly, too, more broadly and vaguely, harmful content.[3] Attention for platforms' powerful intermediary role in online speech precedes 2020, but the pressure on them to 'clean up' their services is unprecedented.

In this chapter[4], we take a close look at how online platforms have responded to health and political disinformation in 2020. We publish detailed comparative timelines of Facebook, Google, TikTok and Twitter's responses to COVID-19 and US general election-related disinformation, in an effort to further our understanding of their content moderation practices. We start by providing a brief sketch of the policy and theoretical context in which these platform responses take place. The editorial role of platforms has become undeniable but is currently largely unregulated. We also explain our methodology and provide details on the dataset we are making publicly available, followed by our comparative analysis of responses by four platforms to the global disinfodemic in 2020.

We conclude that online platforms heavily emphasized amplification of credible COVID-19 related information of the World Health Organization (WHO) and other public health authorities, including through provision of free advertising space. Platforms even launched their own content initiatives, most prominently visible through information panels, but Facebook also livestreamed interviews with leading health professionals and TikTok co-produced media and information literacy videos.

---

2 Julie Posetti and Kalina Bontcheva, 'Disinfodemic: Deciphering COVID-19 Disinformation. Policy Brief 1' (Paris: United Nations Educational, Scientific and Cultural Organization, 2020); Kalina Bontcheva et al., 'Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression' (Geneva and Paris: International Telecommunication Union and United Nations Educational, Scientific and Cultural Organization, 2020).

3 David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (New York: Columbia Global Reports, 2019).

4 A short version of this chapter appeared on the EU Disinfolab blog in February 2021, Trisha Meyer and Alexandre Alaphilippe, 'One Year Onward: Platform Responses to COVID-19 and US Elections Disinformation in Review', 24 February 2021, https://www.disinfo.eu/publications/one-year-onward-platform-responses-to-covid-19-and-us-elections-disinformation-in-review/.

Online platforms also rapidly and regularly expanded their policies, especially on Misleading and Harmful Content, in order to ban, remove, demote or label disinformation harmful but not illegal. Facebook and Google in particular use their advertising policy to aggressively pre-screen paid content on their platforms, and Twitter experimented with additional 'friction',[5] slowing down users' reactions through prompts when users sought to share misleading content during the US elections.

In 2020, the role of online platforms in content moderation became visible as never before. We argue that their ability to react quickly is both encouraging and worrying, if not accompanied by a known hierarchy of principles and stringent transparency and review measures.

2. *Policy and theoretical context*

The legal underpinnings to the current approach to platform regulation find their origins in internet legislation of the late 1990s and early 2000s. Section 230 of the Communication Decency Act (and other sectoral legislation, such as the Digital Millennium Copyright Act) in the United States and the E-Commerce Directive in the European Union were among the first laws granting internet intermediaries limited liability when content generated by their users infringed local intellectual property, speech or security laws.[6] Online platforms did not exist in the same shape or on the same scale as they do now, but the general recognition was that the internet would not be able to grow and flourish if internet intermediaries, whatever shape they took, had to worry all that much about how users were using their services.[7] Only when illegal content and behaviour come to the attention of intermediaries does action need to be taken to remove

5 Ezra Klein, 'The Case for Slowing Everything Down A Bit', *Vox*, 19 November 2018, https://www.vox.com/technology/2018/11/19/18101274/google-alphabet-facebook-twitter-addiction-speed.

6 European Parliament and Council of the European Union, 'Directive 2000/31/EC of the European Parliament and of the Council of the European Union on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market' (2000); US Copyright Office, 'Digital Millennium Copyright Act. Pub. L. No. 105-304, 112 Stat. 2860' (1998); US Congress, 'Communication Decency Act (Title V of the Telecommunications Act)' (1996).

7 Ian Brown and Christopher T. Marsden, *Regulating Code: Good Governance and Better Regulation in the Information Age* (Cambridge, Mass: MIT Press, 2013); Roger Brownsword, *Rights, Regulation, and the Technological Revolution* (Oxford and New York: Oxford University Press, 2008).

egregious content. However, it is clear that this provides a strong incentive to be content-agnostic and play the ignorance card.[8]

This early approach of limited liability fit into a context of a tech-optimism that the internet would bring positive and empowering societal change. It is also a regulatory mirror of the internet's main architectural end-to-end principle to keep the core as efficient and flexible as possible.[9] However, by 2013, the mood towards internet intermediaries started to shift, most notably with Snowden's revelations of mass government surveillance facilitated by telecoms companies.[10] Then, in 2016, fear of undue (foreign) influence in the US general elections and the UK Brexit referendum fanned the flames further. In 2018, the turn towards tech-pessimism was complete when the Facebook-Cambridge Analytica scandal finally came to light. The temptation to harvest user data and getting in front of the influence curve, whether for political or economic gain, proved greater than the early internet rules could curtail.[11]

The result has been a flurry of regulatory inquiries and proposals to curb the excesses of platform power.[12] Some focus on platforms' economic power and consider vigorous application of competition law or new tax rules the way forward; others target the political power gained through micro-targeting and advertising; some still recognize the need to support

8 Heidi Tworek, 'Social Media Platforms and the Upside of Ignorance', *Centre for International Governance Innovation*, 9 September 2019, https://www.cigionline.org/articles/social-media-platforms-and-upside-ignorance.

9 Janet Abbate, *Inventing the Internet* (London & Cambridge, MA: MIT Press, 1999); Brian Carpenter, 'RFC1958. Architectural Principles of the Internet' (Online: Internet Architecture Board, June 1996), https://tools.ietf.org/html/rfc1958.

10 Zygmunt Bauman et al., 'After Snowden: Rethinking the Impact of Surveillance', *International Political Sociology* 8, no. 2 (1 June 2014): 121–44, https://doi.org/10.1111/ips.12048; Julia Pohle and Leo Van Audenhove, 'Post-Snowden Internet Policy: Between Public Outrage, Resistance and Policy Change', *Media and Communication* 5, no. 1 (2017): 1–6, http://dx.doi.org/10.17645/mac.v5i1.932.

11 Robin Mansell, *Imagining the Internet: Communication, Innovation and Governance* (Oxford: Oxford University Press, 2012); Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019).

12 On the rise of platform governance, see Robert Gorwa, 'What Is Platform Governance?', *Information, Communication & Society* 22, no. 6 (12 May 2019): 854–71, https://doi.org/10.1080/1369118X.2019.1573914. For a comparative overview of currently proposed platform regulation, see Freddy Mayhew, 'Regulating Facebook and Google: The Growing Global Big Tech Backlash', *Press Gazette*, 18 February 2021, sec. News, https://www.pressgazette.co.uk/regulating-facebook-google/.

journalism and media and information literacy. Worrying from our perspective is the desire of some regulators to do away with the mostly content-agnostic approach of internet intermediaries. Although the intention to protect vulnerable groups is well-grounded, so was the discomfort felt at the banning of US President Trump by Twitter and Facebook.[13] Private companies decided on their own terms where acceptable speech ends. To be clear, they do this all the time.[14]

Despite this chapter's focus on online platforms, we would like to broaden our horizon momentarily. With the advent of each new technology, we both herald and cower at its invention, but tend to overplay its impact and downplay our agency to chart its course.[15] Importantly, in this chapter we should avoid confusing cause and effect. Recommender systems and algorithms on online platforms exasperate but are not the cause of digital disinformation. In addition, technology ('code' as Lawrence Lessig[16] calls it) is only one of several means of regulating such societal problems. It is powerful but should be considered alongside other approaches (which Lessig divides into law, market, and norms). Platforms are not off the hook, but a comprehensive approach is needed.

Indeed, there is reason for concern at outsourcing speech control to online platforms. This should not be in hands of private corporations, especially when they are largely unaccountable and the explainability of their decision-making leaves much to be desired.[17] The emphasis should therefore not be on expanding content moderation from illegal to harmful content, but rather on creating transparency in the process of content

13 Alex Hern, 'Opinion Divided over Trump's Ban from Social Media', *The Guardian*, 11 January 2021, https://www.theguardian.com/us-news/2021/jan/11/opinion-divided-over-trump-being-banned-from-social-media.

14 Judit Bayer, 'Between Anarchy and Censorship. Public Discourse and the Duties of Social Media', CEPS Paper in Liberty and Security in Europe No. 2019-03 (Online: CEPS, May 2019), https://www.ceps.eu/ceps-publications/between-anarchy-and-censorship/.

15 Andrew Feenberg, *Questioning Technology* (London & New York: Routledge, 1999); Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven: Yale University Press, 2018).

16 Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0* (New York: Basic Books, 2006).

17 Kaye, *Speech Police*. For an academic and civil society discussion on minimum standards for content moderation, see ACLU Foundation of Northern California et al., 'Santa Clara Principles on Transparency and Accountability in Content Moderation', Santa Clara Principles, accessed 15 March 2021, https://santaclaraprinciples.org/images/scp-og.png.

moderation. Their editorial role is evident, but how decisions are made currently is not. We will return to these thoughts at the end of the chapter.

3. *Methodology and dataset*

For this chapter, we reconstructed a timeline of responses of Facebook, Google, TikTok and Twitter, on the basis of reports submitted to the European Commission as part of the Fighting COVID-19 Disinformation Monitoring Programme, as well as updates posted on their company blogs.[18] It is important to note that we analysed what platforms announced and reported, not whether these measures were implemented.

We mapped their responses by month and against the platform disinformation response typology we developed as part of our contribution to the UNESCO/ITU Balancing Act study mentioned in the introduction.[19] In particular, we divide platform responses into four types of 'content' mod-

---

18 We used the sources below to map the platform responses on a month-by-month basis. This was not always a straightforward exercise, and we would be very happy to rectify any error you may spot!
We did not include company updates related to support for health workers, small businesses, non-profits, children, social movements, communities, mental health, emotional well-being or diversity, as these were not specific to combating disinformation on the platforms.
All: monthly platform reports from August 2020 for the European Commission Fighting COVID-19 Disinformation Monitoring Programme, https://ec.europa.eu/digital-single-market/en/news/first-baseline-reports-fighting-covid-19-disinformation-monitoring-programme
Facebook: Facebook Coronavirus Newsroom updates, https://about.fb.com/news/2020/12/coronavirus/; Facebook US 2020 Elections report, https://about.fb.com/actions/preparing-for-elections-on-facebook/; Facebook Key Elections Investments and Improvements timeline, https://about.fb.com/wp-content/uploads/2020/12/Elections-Investments-and-Improvements.pdf
Google: Google Keyword COVID-19 updates, https://blog.google/inside-google/covid-19/; Elections Google updates, https://elections.google/ - updates
TikTok: TikTok Safety Center – COVID-19, https://www.tiktok.com/safety/resources/covid-19?lang=en&appLaunch=; TikTok Safety updates, https://newsroom.tiktok.com/en-us/safety; TikTok Integrity for the US Elections, https://www.tiktok.com/safety/resources/2020-us-elections
Twitter: Twitter Blog, https://blog.twitter.com/; Twitter Coronavirus updates, https://blog.twitter.com/en_us/topics/company/2020/covid-19.html; Twitter Blog Elections tag, https://blog.twitter.com/en_us/tags.blog–elections.html

19 Bontcheva et al., 'Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression'.

eration: flagging/labelling, blocking/removing, limiting/demoting and prioritizing/amplifying; and four types of 'other' moderation: specific to accounts, advertising, users, and research/review.

The aim of this breakdown into different types of 'moderation responses' is to map online platforms' change in emphasis over time in a granular fashion. Each is a different manifestation of the editorial role that platforms play in moderating online speech.

Table 1 below shows our mapping for Facebook's responses to COVID-19 and US election disinformation in 2020 as an illustration (this is only 1/8th of the dataset). We cordialy invite you to consult the complete dataset online[20]. In this online resource, we publish two timelines, with the data organized by platform and by response type.

20 Trisha Meyer, 'Comparative Timeline of Platform Responses to COVID-19 and US Elections Disinformation, Organised by Platform and by Response (Updated Regularly)', Google Sheets, 18 February 2021, https://bit.ly/3ySbJXc.

Table 1. Facebook responses to COVID-19 and US elections related disinformation in 2020 (own compilation)

FACEBOOK	Continuous	Jan	Feb	March	April	May	June	July	Aug	Sep	Oct	Nov	Dec
Flagging / labelling content	*gradually expands fact-checking collaborations; increases use of AI to detect misinformation and deep fakes*			commits to invest $2M additional funding for fact-checking, news and research on misinformation; content moderators are sent home	provides educational pop-ups for users who reacted to harmful COVID-19 misinformation; content moderators partially return to office		starts labelling of state-controlled media	starts adding labels to posts on voting, including federal officials and candidates		launches second wave of funding for fact-checking and news; announces stronger labels on posts from candidates and campaigns that prematurely claim victory or attempts to delegitimize the elections	promotes awareness campaigns of public health agencies on vaccines		

Blocking / removing content	announces policy on removal of misleading manipulated media	removes COVID-19 disinformation "with imminent physical harm" on Facebook and Instagram	removes COVID-19 information from recommendations "unless posted by a credible health organization" (Instagram)			announces stricter policy on content that interferes with or suppresses voting		removes health groups from recommendations; announces stronger voter restriction policies	bans calls for poll watching when combined with militarized language		removes debunked false claims on vaccines
Limiting / demoting content			demotes and issues warnings on debunked COVID-19 disinformation that falls within community guidelines			demotes and issues warnings on debunked voting disinformation that falls within community guidelines		sets forwarding limit on Messenger		redirects searches for terms related to Qanon on Facebook and Instagram to credible resources of GNET	
Prioritizing / am-		displays educational	launches COVID-19 Edu-	expands educational		launches Voting Informa-	global reminder to wear				

pop-ups of WHO and public health officials in search on Facebook and Instagram	cational Centre, in collaboration with WHO, on news feed (Facebook); shows information of WHO and public health agencies at top of feed and adds stickers to promote accurate information (Instagram); launches information hub and WHO Health Alert	pop-ups to Groups, prompts group admins to share live broadcasts from public health officials; partners with DCD to develop COVID-19 curriculum for group admins	tion Centre; starts prioritizing original news reporting	face coverings, adds facts about COVID-19 to information centre	

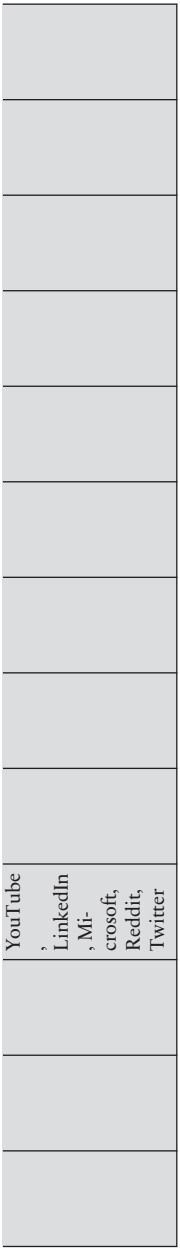
			(WhatsApp); launches Messenger Coronavirus Community Hub (Messenger)							
					starts verifying identity of people with high-reach posts on Facebook in the USA					
Account-specific responses	*continues action and collaboration against Coordinated Inauthentic Activity*						expands Dangerous Individuals and Organizations Policy to combat groups tied to violence (such as QAnon)	reinforces User Recidivism Policy		updates enforcement against QAnon, removing any Facebook pages, groups and Instagram accounts representing Qanon, even if they contain no violent content

Ad-specific responses	*continues use of Ads Library*	expands required authentication of advertisers for political/social issue ads	bans ads promoting COVID-19 cure	provides free advertising space to WHO and public health agencies for awareness raising on COVID-19; temporarily bans advertisements and commerce listings for COVID-19 related products			adds US House and Senate ad tracker to Ad Library; announces stricter policy on ads that interfere with or suppress voting; allows promotion and trade in non-medical masks		allows promotion and sale of hand sanitizers and surface disinfection wipes	announces restrictions on new political/social issue ads in final week of campaign; announces ban on ads that prematurely claim victory or attempt to delegitimize the elections	bans ads discouraging vaccines; announces temporary suspension of social issue, electoral and political ads after polls close	blocks creation of new ads about social issues, elections or politics immediately before election day, temporarily suspends ads about social issues, elections or politics after election day	updates ad policy for COVID-19 vaccines, e.g. allowing ads on safe access to vaccine, but prohibiting sale or expedited access
User-specific responses				launches digital literacy programme (Get Digital); live interview between Mr Zuckerberg, Ms Chan			adds new control that allows people to see fewer political/		live interview between Mr Zuckerberg and Dr Fauci			live interview between Mr Zuckerberg and Dr Fauci	

		view between Mr Zuckerberg and Dr Fauci	and Dr Frieden		social issue ads on Facebook and Instagram; launches media literacy campaign (Stamp Out False News)				
Review / research-specific responses		connects health organisations and developer partners to scale response to COVID-19; supports COVID-19 developer hackathon; content	announces new Data for Good tools to help health researchers track and combat COVID-19; content moderators partially re-		releases new visualisations, data sets and new survey to combat COVID-19		launches Presidential Election Research Initiative		

			moderators are sent home, increases reliance on automated technology	turn to office					
		launches Election Operations Centre for primary elections in the USA	commits to invest $100M in news industry; $1M through local news covering COVID-19 in USA and Canada; joint announcement on COVID-19 coordination from Facebook, Google,						
increases use of AI to detect misinformation and deep fakes									

Other



4. Platform-specific responses

In this section, we highlight Facebook, Google, TikTok and Twitter's main responses to COVID-19 and US general election-related disinformation in 2020 as a basis for our comparison.

Facebook (Facebook, Messenger, Instagram, Whatsapp)

As our timeline shows, Facebook was busy, with a frenzy of activity in February and March 2020 as the COVID-19 pandemic broke out; a steep ramping up of US election-related activities as of June; and a gradual response in preparation for the COVID-19 vaccine rollout as of September. Their COVID-19 response emphasizes the *prioritization* of authoritative content, *free advertising* for public health agencies and *demotion* of debunked information. They also remove COVID-19 related disinformation with 'imminent physical harm'. Meanwhile Facebook's US election response focused on policies related to *political and social issues ads* and policies that allow for *removal* of content that interfered with or suppressed voting. They also added *warning messages* to debunked content.

Google (Search, YouTube, AdSense)

Google's COVID-19 response was gradual, starting with *prioritization* and amplification of accurate COVID-19 related content and free *advertising* credits for public health authorities. Notably they also published a COVID-19 Medical Misinformation policy and expanded their Harmful Health Claims policy to *remove* content that contradicts authoritative and scientific consensus on the health crisis. Google's US election response focused on *security* and *amplification* of trusted news. It is important to note that *advertisement*-related policies are a powerful tool for Facebook and Google to wield. Both Facebook and Google temporarily paused US election ads after the polls closed.

TikTok

TikTok's COVID-19 response started earlier than other platforms and was concentrated in time (Jan-March). A similar approach was followed for

vaccines in December. It stresses information *prioritization* and amplification through in-app notices, stickers, and brand takeovers. In October TikTok launched Project Halo, a science *communication* effort, to raise awareness and confidence in vaccine. TikTok's US election response was similarly concentrated in time (Aug-Oct) and focused on an in-app guide and *public service announcements*. During the month of October until the end of Election Day, they provided daily updates on their election response. TikTok does not allow political ads. Similar to Facebook and Google, it donated *ad space* to public health authorities. In February 2021, TikTok announced that they will add *friction* to their disinformation response arsenal. When they identify a video with unsubstantiated claims, Tiktok will show a banner *warning* and include several warning prompts before viewers share a flagged video.

Twitter

In 2020 Twitter played an extensive editorial role on its platform, through use of *labels*, *warnings*, *removal*, *reducing visibility*, *adding friction*, *promoting authoritative content*. As part of its COVID-19 response, Twitter broadened its policy definition of harm to include content that contradicts COVID-19 public health guidance. In February and May, they also issued guidance on their staged approach to manipulated and synthetic media and potentially misleading content. A frenzy of activity occurred in the lead up to and aftermath of US elections on *content and account* level. In December, Twitter reported that their more extensive version of friction (Quote Tweet rather than Retweet; removing 'liked by' and 'followed by' recommendations, only surfacing 'additional context' trends) did not bear expected results.

5. Comparison and key take-aways

Table 2. Comparison of platform responses to COVID-19 and US election-related disinformation (own compilation) [main responses related to C = COVID-19; E = US elections]

Main response type per platform	Facebook	Google	TikTok	Twitter
Flagging/labelling content	E		C	C / E
Blocking/removing content	C/ E	C	C / E	C / E

Limiting/demoting content	C / E		E	E
Prioritizing/amplifying content	C	C / E	C / E	C / E
Account-specific	E			E
Advertising-specific	C / E	C / E	C	C
User-specific			C	
Review/disinformation research-specific				

The timing of online platform responses to COVID-19 corresponds with the arrival of the virus in Europe and North America in March 2020 – despite having users globally. TikTok, the only non-Western (Chinese) social media company in our sample, is an exception. Its COVID-19 response ramps up in February 2020. Similarly striking is the platforms' response to US election disinformation. Platforms have come a long way since the 2016 US general elections and UK Brexit referendum. They have been prompted by governments to keep records of political ad spending, to mitigate foreign interference, to ensure fair and free elections. Yet the actions taken by platforms in the 2020 US elections were unprecedented. In particular, in a span of only a few months, labelling and removing political speech and figures became normalized.

As the pandemic hit, we saw online platforms heavily *prioritize authoritative content* provided by public health officials through in-app notices, educational pop-ups and prompts, launching dedicated hashtags and educational centres, and surfacing credible public health information at the top of feeds and in COVID-19 related searches. Six months later, similar action to emphasize authoritative content was taken in preparation for the US elections, and towards the end of 2020 to counter vaccine disinformation. One relatively novel development were the grants of *free advertising credits* to the World Health Organization (WHO) and public health authorities. Google, Facebook and Twitter also provided large grants for journalism and fact-checking.

In parallel – and prominently used during the run-up to and in the aftermath of the US general elections – platforms regularly updated their *policies* related to Misleading and Harmful Content, Sensitive Events, Civic Integrity to *ban, remove or demote content and ads* that contradicted public COVID-19 health guidance and undermined confidence in the elections. Efforts to counter QAnon led Facebook to expand its Dangerous Individuals and Organizations Policy to include organizations tied to violence in August 2020. Much later, in January 2021, Twitter updated its Coordi-

nated Harmful Activity Policy. Infamously, both platforms permanently suspended President Trump's accounts in January 2021 for inciting the violence at the Capitol Hill riots.

In 2020, platforms also extended their use of warning messages and stickers to *label and flag* potentially misleading content, caution to share further and point to credible information. In February 2020, Twitter started taking action against synthetic and manipulated media, and in May against potentially misleading COVID-19, election- (and vaccine-) related content (see Figure 1 below). Slightly later, in June 2020, Facebook started labelling state-controlled media, and in July the accounts of political candidates and federal officials.

Figure 1: Twitter's approach to misleading content (May 2020)[21]

Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
	Propensity for Harm	

6. *Conclusion*

In response to game changer events such as the global health pandemic or the use of disinformation by US President Trump, online platforms took unprecedented measures to minimize harm by improving their content moderation efforts. Some policy updates were clearly planned, such as Twitter's graduated response to synthetic and manipulated media, while others were kneejerk responses to ongoing events. This rapid expansion of

21 Yoel Roth and Nick Pickles, 'Updating Our Approach to Misleading Information', *Twitter Blog* (blog), 11 May 2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

disinformation policies culminated in bans of President Trump and Parler on multiple platforms in January 2021.

This ability to react quickly is both encouraging and worrying. Encouraging, because it demonstrates that platforms, under societal pressure, behave as a public interest utility in specific cases. Yet at the same time, it is worrying as the majority of measures taken fail to address the root causes of the architecture of information distribution. Without this, discussions around censorship and its abuses will prevail over the work needed to build a more inclusive information ecosystem.

The emphasis of current regulatory discussions on platforms needs to be on accountability. Our analysis was based on information we were able to gather from company blogs and reports submitted to the European Commission. While their reporting is a step forward, this information only became comparable data with significant additional effort but it does not offer insights into the implementation or consequences of action taken.

We need detailed metrics on online content distribution. Crucially this should include transparency in terms of content promotion/demotion in addition to removal of content, as an initial means of auditing algorithms. The online advertising ecosystem is also deserving of reinforced scrutiny to gain a better understanding of the impact of changes in ad policies. In light of platforms' emphasis on granting advertising credit, it seems appropriate to establish a register of beneficiaries of ad-credits detailing amounts granted and spent. Civil society (academics, researchers, journalists, civil society organizations) and independent regulators should also be empowered in their role of enforcing accountability of online platforms. The stick behind the door might need to be available to sanction bad faith actors, especially when there are repeated efforts to escape transparency and accountability.

Finally, to return to the opening paragraph of the chapter, regulating platforms will be in vain if we do not tackle the causes of the disinfodemic at the same time. This requires rebuilding trust by listening to others, celebrating our differences, and committing to common objectives. If 2020 can teach us anything, we hope it is that our social bonds and communities are more resilient than we perhaps thought yet also require continual collective and individual commitment.

Bibliography

Abbate, Janet. *Inventing the Internet*. London & Cambridge, MA: MIT Press, 1999.

ACLU Foundation of Northern California, Center for Democracy & Technology, Electronic Frontier Foundation, New America's Open Technology Institute, Irina Raicu, Nicolas Tuzor, Sarah Myers West, and Sarah T. Roberts. 'Santa Clara Principles on Transparency and Accountability in Content Moderation'. Santa Clara Principles. Accessed 15 March 2021. https://santaclaraprinciples.org/images/scp-og.png.

Bauman, Zygmunt, Didier Bigo, Paulo Esteves, Elspeth Guild, Vivienne Jabri, David Lyon, and R. B. J. Walker. 'After Snowden: Rethinking the Impact of Surveillance'. *International Political Sociology* 8, no. 2 (1 June 2014): 121–44. https://doi.org/10.1111/ips.12048.

Bayer, Judit. 'Between Anarchy and Censorship. Public Discourse and the Duties of Social Media'. CEPS Paper in Liberty and Security in Europe No. 2019-03. Online: CEPS, May 2019. https://www.ceps.eu/ceps-publications/between-anarchy-and-censorship/.

Bontcheva, Kalina, Julie Posetti, Denis Teyssou, Trisha Meyer, Sam Gregory, Clara Hanot, and Diana Maynard. 'Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression'. Geneva and Paris: International Telecommunication Union and United Nations Educational, Scientific and Cultural Organization, 2020.

Brown, Ian, and Christopher T. Marsden. *Regulating Code: Good Governance and Better Regulation in the Information Age*. Cambridge, Mass: MIT Press, 2013.

Brownsword, Roger. *Rights, Regulation, and the Technological Revolution*. Oxford and New York: Oxford University Press, 2008.

Carpenter, Brian. 'RFC1958. Architectural Principles of the Internet'. Online: Internet Architecture Board, June 1996. https://tools.ietf.org/html/rfc1958.

European Parliament and Council of the European Union. Directive 2000/31/EC of the European Parliament and of the Council of the European Union on Certain Legal Aspects of Information Society Services, in particular Electronic Commerce, in the Internal Market (2000).

Feenberg, Andrew. *Questioning Technology*. London & New York: Routledge, 1999.

Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press, 2018.

Gorwa, Robert. 'What Is Platform Governance?' *Information, Communication & Society* 22, no. 6 (12 May 2019): 854–71. https://doi.org/10.1080/1369118X.2019.1573914.

Hern, Alex. 'Opinion Divided over Trump's Ban from Social Media'. *The Guardian*, 11 January 2021. https://www.theguardian.com/us-news/2021/jan/11/opinion-divided-over-trump-being-banned-from-social-media.

Kaye, David. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports, 2019.

Klein, Ezra. 'The Case for Slowing Everything Down A Bit'. *Vox*, 19 November 2018. https://www.vox.com/technology/2018/11/19/18101274/google-alphabet-facebook-twitter-addiction-speed.

Lessig, Lawrence. *Code: And Other Laws of Cyberspace, Version 2.0*. New York: Basic Books, 2006.

Mansell, Robin. *Imagining the Internet: Communication, Innovation and Governance*. Oxford: Oxford University Press, 2012.

Mayhew, Freddy. 'Regulating Facebook and Google: The Growing Global Big Tech Backlash'. *Press Gazette*, 18 February 2021, sec. News. https://www.pressgazette.co.uk/regulating-facebook-google/.

Meyer, Trisha. 'Comparative Timeline of Platform Responses to COVID-19 and US Elections Disinformation, Organised by Platform and by Response (Updated Regularly)'. Google Sheets, 18 February 2021. https://docs.google.com/spreadsheets/d/1KR-YECATOyEHy_jd1pWXsujuHeL8sE6WWZpk7Ib8LSM/edit?usp=sharing&usp=embed_facebook.

Meyer, Trisha, and Alexandre Alaphilippe. 'One Year Onward: Platform Responses to COVID-19 and US Elections Disinformation in Review', 24 February 2021. https://www.disinfo.eu/publications/one-year-onward-platform-responses-to-covid-19-and-us-elections-disinformation-in-review/.

Pohle, Julia, and Leo Van Audenhove. 'Post-Snowden Internet Policy: Between Public Outrage, Resistance and Policy Change'. *Media and Communication* 5, no. 1 (2017): 1–6. http://dx.doi.org/10.17645/mac.v5i1.932.

Posetti, Julie, and Kalina Bontcheva. 'Disinfodemic: Deciphering COVID-19 Disinformation. Policy Brief 1'. Paris: United Nations Educational, Scientific and Cultural Organization, 2020.

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

Sacks, Jonathan. *Morality: Restoring the Common Good in Divided Times*. New York: Basic Books, 2020.

Tworek, Heidi. 'Social Media Platforms and the Upside of Ignorance'. *Centre for International Governance Innovation*, 9 September 2019. https://www.cigionline.org/articles/social-media-platforms-and-upside-ignorance.

US Congress. Communication Decency Act (Title V of the Telecommunications Act) (1996).

US Copyright Office. Digital Millennium Copyright Act. Pub. L. No. 105-304, 112 Stat. 2860 (1998).

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.