6 Responses within Production and Distribution

Chapter 6 of the report:

Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet'

Published in 2020 by International Telecommunication Union (ITU), Place des Nations, CH-1211 Geneva 20, Switzerland, and the United Nations Educational, Scientific and Cultural Organization, and United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, Place de Fontenoy, 75352 Paris 07 SP, France

ISBN 978-92-3-100403-2

This research will be available in Open Access under the Attribution-ShareAlike 3.0 IGO (CC-BY SA 3.0 IGO) license. By using the content of this publication, the users accept to be bound by the terms of use of the UNESCO Open Access Repository

https://en.unesco.org/publications/balanceact

6.1 Curatorial responses

Authors: Trisha Meyer, Clara Hanot and Julie Posetti

This chapter discusses measures to tackle disinformation through content curation or moderation within internet communication companies and journalism processes. The effect of such measures affects inter alia what content is allowed on the service; if it is allowed to remain up, if it is fact-checked; its prominence and visibility; whether advertising appears next to it; the degree to which it is automatically recommended or limited in terms of distribution; whether it is labelled, as well as what kinds of paid content appear and how. The issues involved relate to the policy provisions, their enforcement, and the issue of redress. These implicate all online content including information and disinformation.

Curatorial responses within the internet companies are primarily addressed via their policies, which we analyse in this chapter. These responses often result in technical or algorithmic measures, which are covered in depth in Chapter 6.2 Technical/Algorithmic responses. These responses also involve normative and ethical elements, which are addressed in chapter 7.1.

News organisations, journalists and other publishers of public interest information also respond curatorially to the problem of disinformation. Such functions can include reporting based on collaborative fact-checking, editorial curation of knowledge and resources, collaborative fact-checking partnerships, curation of sources and resources, audience curation (e.g. User Generated Content), and comment moderation. Chapters 4.1, 4.2 and 5.3 deal with editorial curation efforts associated with fact-checking and investigative reporting. Ethical and normative issues associated with editorial curation are addressed in chapter 7.1, and training initiatives related to curation of disinformation within media institutions are addressed in Chapter 7.2 which deals with educational responses.

Below, the terms of service, community guidelines and editorial policies of 11 internet communications companies (Facebook, Instagram[181], WhatsApp[182], Google, YouTube[183], Twitter, VK, Weibo, WeChat, LINE and Snapchat) are examined to gain an in-depth understanding of how these companies expressly or indirectly address the problem of disinformation. These actions tend to seek to curb manipulative actors, deceptive behaviours, and what is perceived to be potentially harmful content (François, 2019). Attention is paid to how decisions on content curation/moderation are made, whether/how users or third parties are enlisted to help with content monitoring, and which appeal mechanisms are available.

Actions undertaken by these companies may be efficient and dynamic, but questions are also raised by various actors regarding the regulatory purview granted through this process to private commercial actors. Concerns about the somewhat random application of self-regulatory measures - for example, emphasising responses in the U.S. environment while abrogating responsibilities in high risk countries in the Global South (Ingram, 2018)

[181] Note: Instagram is owned by Facebook
[182] Note: WhatsApp is owned by Facebook
[183] Note: YouTube is owned by Google

- and the limiting of measures due to the prioritisation of profit, have led to calls for self regulation to be overridden via independent regulatory mechanisms. The COVID-19 crisis of ubiquitous disinformation has further amplified concerns whether internet communications companies can address the problem through stronger self-regulatory curatorial actions (McNamee, 2020).

Another example is Facebook's controversial policy which exempts categories of political advertising from fact-checking (see chapters 4.1 and 7.1 for discussion of Facebook's fact-checking policy and exemptions). This was critiqued by the company's former 'head of global elections integrity ops' in an article titled *I worked on political ads at Facebook. They profit by manipulating us* published by the Washington Post. Yael Eisenstat (2019) wrote "The real problem is that Facebook profits partly by amplifying lies and selling dangerous targeting tools that allow political operatives to engage in a new level of information warfare." More recently, the *Wall Street Journal* published leaked detail from a Facebook team presentation which warned of the risks of the company's algorithmic curation: "Our algorithms exploit the human brain's attraction to divisiveness...If left unchecked [users would be fed] more & more divisive content in an effort to gain user attention & increase time on the platform" (Horowitz & Seetharaman, 2020). Facebook responded, saying: "If Pages and Groups repeatedly share content that violates our Community Standards, or is rated false by fact-checkers, we reduce their distribution, remove them from recommendations, and we remove the ability for Pages to monetize and advertise. We also remove entire Pages and Groups who repeatedly post violating content." (Rosen, 2020)[184] Before the decision by YouTube to 'deplatform' conspiracy-monger Alex Jones, the company's algorithm was said by a former employee to have recommended his "info-wars" videos more than 15,000,000,000 times.[185]

Built to locate content and/or connect users, facilitate the curation and sharing of content, and seeding engagement with it, the main features of internet communications companies can be exploited to spread disinformation. That is, tools that initially allowed freedom of expression and access to information to flourish have been weaponised against truth, accuracy and access to credible public interest information (Posetti et al., 2019a). A typical strategy adopted by disinformation agents to share false or misleading content involves attaching a catchy headline to an emotionally provocative 'story' (which either does not fulfil the promise of the headline, or is based on fabricated information) to drive engagement and clicks. This is known as clickbait. It has been demonstrated that emotionally-charged content tends to generate higher interactions (Martens et al., 2018). Attracting engagement, likes, and shares, deceptive actors can take advantage of the network effect provided by the platforms' algorithms tailored to surface relevant content to the user, thus accelerating and broadening reach for deceitful messages (DiResta, 2018). Measures taken by internet communications companies towards reducing clickbait are discussed in Chapter 6.2.

In a digital advertising economy, these companies act as de facto 'attention brokers' (Wu, 2017). They have to strike a difficult balance given that the content with the most engagement is also the most lucrative in terms of data collection and/or associated advertising delivery. Data breaches (e.g. Cambridge Analytica), foreign interference in democratic elections (e.g. US 2016; Howard et al., 2018), and the massive diffusion of disinformation via messaging apps in the context of the elections (e.g. India 2019) and

184 However, as noted throughout this report, at the time of writing Facebook was continuing to exclude political advertising and content posted by politicians from such counter disinformation measures.

185 https://twitter.com/gchaslot/status/967585220001058816?s=21

health crises, such as the pandemic associated with COVID-19 (see specific discussion below), have put pressure on the companies to take actions to mitigate the propagation of disinformation content on their services (Burgos, 2019).

Much communications online relies on intermediaries, and in the process is mediated by policies. These may include human fact-checkers, moderators and investigators, including those employed by news organisations, internet communication companies, as well as partnerships with news organisations and other verification experts[186]. Such communications are also mediated via digital architecture – the technical protocols that enable, constrain, and shape user behaviour online, and which reflect business models and other considerations. These technical and design features differ from one service to another. Concretely, how connections between accounts on social media are created and maintained, how users can engage with each other via the technology, as well as the algorithmic filtering and datafication, all shape the way communication (as well as search and discovery) is tailored on a specific platform (Bossetta, 2018). These variations in purpose and architectural structure also partly explain why curation strategies can differ in some respects from one company to another.[187] These tensions and challenges of using curatorial responses to support or defend freedom of expression are further elaborated in the evaluation in the last section of this chapter.

Key to successful curatorial responses is independent oversight. In this context, civil society organisations and citizens play an important role, since they can continuously check the ways in which social platforms protect freedom of expression and implement full transparency in their curatorial actions. Transparency, accountability, and appeal in relation to curatorial actions are essential for protecting freedom of expression, and necessary since the platforms' algorithms and moderators do make mistakes. Given the vast numbers of users and daily posts on the platform, if left unchecked these curational impacts can amount to a significant problem.

6.1.1 Internet communication companies' approaches to content curation

This section provides an overview of how internet communication companies curate or moderate content and accounts based on their terms of service, community guidelines and editorial policies.[188]

Below is the list of primary sources used in the analysis of each platform:

Facebook and Instagram

https://www.facebook.com/communitystandards/introduction ; https://help.instagram.com/477434105621119 ; https://transparency.facebook.com/ ; Facebook & Instagram (2019)

186 See chapters 4.1 (monitoring and fact-checking) and 5.3 (electoral responses) for a detailed discussion of the curatorial role of fact-checking

187 As an example, during the 2018 Irish referendum on the Thirty-sixth Amendment of the Constitution Act (on abortion), Google decided not to accept political advertising, whereas Facebook only banned foreign actors' adverts. Based on its advertising policy, Twitter banned abortion adverts from outset (O'Brien & Kelly, 2018; Satariano, 2018).

188 We are grateful to our fellow researchers who took precious time to read and analyse the terms of service, community guidelines and editorial policies of Weibo and WeChat (Olivia Sie), VK (Vsevolod Samokhvalov) and LINE (Koji Yamauchi) in the platforms' predominant user language (Chinese, Russian, Japanese).

WhatsApp

https://www.whatsapp.com/legal/?eea=0#terms-of-service ; https://faq.whatsapp.com/21197244/#Report ; https://blog.whatsapp.com/10000647/More-changes-to-forwarding ; WhatsApp (2019)

Google and YouTube

https://about.google/community-guidelines/ ; https://transparencyreport.google.com ; https://www.youtube.com/yt/about/policies/#community-guidelines ; Google and YouTube (2019)

Twitter

https://help.twitter.com/en/rules-and-policies#research-and-experiments ; https://transparency.twitter.com/en/removal-requests.html ; Twitter (2018) ; Twitter (2019)

VK

https://vk.com/licence ; https://vk.com/blog ; https://vk.com/support?act=home ; https://vk.com/help?page=cc_terms ; https://vk.com/page-76477496_50995734

Weibo

https://www.weibo.com/signup/v5/protocol

WeChat

https://www.wechat.com/en/service_terms.html (international users) ; https://weixin.qq.com/cgi-bin/readtemplate?lang=en&t=weixin_agreement&s=default&cc=CN (mainland China users) ; https://help.wechat.com/ ; https://wechatwiki.com/wechat-resources/wechat-rules-and-marketing-restrictions/

LINE

https://terms.line.me/line_terms/?lang=ja ; LINE (2020)

Snapchat

https://www.snap.com/en-US/community-guidelines ; https://www.snap.com/en-US/ad-policies/political/ ; https://businesshelp.snapchat.com/en-US/article/political-ads-library

The focus is on internet communication companies (social media, messaging, video sharing, search), as they have been at the centre of requests to tackle disinformation online. In reviewing their terms of service, community guidelines and editorial policies ('platform rules'), the following curatorial responses and dimensions can be discerned:

1. Flagging and review of content
2. Filtering, limiting, blocking or removal of content
3. Promotion/demotion of content
4. Disabling or removal of accounts
5. Transparency in sponsored content
6. User involvement
7. Appeal mechanisms

In the table below, the actions taken by 11 geographically diverse and global companies that enjoy a large user base are mapped. In the subsequent analysis, differences in the curation of content and accounts between these companies are detailed, with examples provided. The analysis is based on documentation (policies, blogs, transparency reports) pertaining to content curation, provided by the internet communications companies. The table only marks actions for which evidence was found in the documentation. Where no (or insufficient) evidence was found, the action was left blank. If an action is marked between brackets, this signifies that action is dependent on the type of content or user.

Content / account moderation		Facebook Instagram	WhatsApp	Google YouTube	Twitter	VK	Weibo	WeChat	LINE	Snap chat
Flagging and review of content	Machine driven	x		x	x	x	x	x	x	
	Human driven	x		x	x	x	x	x	x	x
	Third party review	x	(x)	x	x	x				
	External counsel	x		x						
Filtering, limiting, blocking and removal of content	(Re-)upload filter	x		x	x	x	x	x		
	Restricted content forwarding		x							
	Restrictions based on:									
	• company rules	x	x	x	x	x	x	x	x	(x)
	• law enforcement	x	x	x	x	x	x	x	x	x
Promotion and demotion of content	Promotion of authoritative sources	x		x	x	x				(x)
	Demotion of clickbait or contested content	x		x	x	x		x		
Disabling and suspension of accounts	Graduated approach:									
	• warning			x						
	• limited features	x	x	x	x		x	x	x	
	• suspension	x	x	x	x	x	x	x	x	x
Transparency in sponsored content	Demarcation of sponsored content	x		x	x	x	x	x	x	x
	Ad transparency centre	x		x	x					x

User involvement	User can flag content for review	x	x	x	x	x	x	x	x	
	User can block/snooze content/ accounts	x	x	x	x	x	x	x	x	x
	User can prioritise content/ accounts	x		x	x	x	x	x	x	x
	User can change advertising categories s/he is placed in	x		x	x				x	
Appeal	Notice of action	x		x	x			(x)	x	
	Possibility to appeal	(x)		(x)	(x)	(x)			x	
	Notification of appeal decision	(x)		(x)	(x)	(x)				

Table 5. *Curatorial responses from internet communication companies*

What do these companies' terms of service, community guidelines and editorial policies actually contain? They provide detail on which type of content prompts action - ranging from violent and criminal behaviour (violence and incitement, individuals and organisations deemed terrorist or criminal, promoting or publicising crime, coordinating harm, violations for regulated goods, fraud and deception, election interference) to objectionable content (hate speech, violence and graphic content, adult nudity and sexual activity, sexual solicitation, cruel and insensitive treatment, bullying), and more.

1. Flagging and review of content

Potentially abusive or illegal content on online communication platforms can be flagged through automated machine learning, and manually by users and third party organisations (e.g. law enforcement, fact-checking organisations, news organisations operating in partnership). Automated detection is on the rise and is important to tackle concerted efforts of spreading disinformation, along with other types of communications deemed potentially harmful (see next Chapter 6.2 Technical/Algorithmic responses). To illustrate the automation of content moderation, over the period from July to September 2019, a total of 8,765,893 videos were removed from **YouTube**. Of these, only 602,826 were reported by humans.[189] On human detection, **Twitter**, for instance, boasts a Partner Support Portal, a fastlane for Twitter partners to receive responses to reports and facilitate information sharing (Twitter, 2019). Other platforms have similar privileged partners, especially law enforcement authorities, with whom they collaborate.

Most online platforms employ staff to review content. Facebook and Google in particular have increased their content moderation staff over the years. **Facebook** employed 15,000 staff to deal with content review in 2019 (Facebook, 2019), while **Google** announced in 2017 that it would hire 10,000 content moderators by the end of 2018 (Hoggins, 2019). **Twitter**'s content moderation staff comprised about 1,500 in 2019. The majority of online platforms' content moderators work as external contractors (Dwoskin, Whalen & Cabato, 2019). At **VK**, a team of more than 100 people, divided into several groups based on the characteristics of detected 'violations', has been engaged in the curation of content.

Automated machine learning is also used to detect disinformation and spam.[190] As the COVID-19 pandemic unfolded, most of these companies moved towards heavy use of

189 https://transparencyreport.google.com/youtube-policy/removals
190 https://vk.com/page-76477496_50995734

automation for content curation. The issue of use of automation in content curation is discussed further in the next chapter (6.2) on algorithmic and technical responses.

In other countries, fact-checking organisations have set up their own accounts to flag suspected false information for verification. Even though some might be supported by the companies, these services are not directly managed by the internet communications companies and they do range wider than content referred to them by these entities (Tardáguila, 2019).

Finally, **Facebook** and **Google** work with external actors such as legal counsel where necessary to verify whether a piece of content breaches standards and/or (especially national) legislation. In 2019 Facebook announced a plan to set up an Oversight Board to "protect free expression by making principled, independent decisions about important pieces of content and by issuing policy advisory opinions on Facebook's content policies" (Facebook, 2019). The first board members were announced in May 2020 (Wong, 2020a). Hailed as a "Supreme Court of content" in some quarters, the initial expectation was that the Board would curtail Facebook's policy on allowing untruths in political advertising (Levy, 2020). However in June 2020, the remit of the Board appeared to be limited to reviewing decisions to remove content (See discussion in chapter 7.1 about normative and ethical responses to disinformation). For its part, **Twitter** has a Trust and Safety Council[191] which provides advice on the application of Twitter's safety rules.

2. Filtering, removal, blocking and other restrictions of content

Interventions that impact on the availability of content are implemented on the basis of the companies' terms of service, community guidelines, editorial policies or law enforcement (see also Chapter 5.1 Legislative, pre-legislative and policy responses). It can be noted that these rules can be more restrictive than their legal basis in a number of jurisdictions. A good example is Twitter's decision to ban paid political advertising globally from in November 2019. At the other end of the spectrum, Facebook decided to continue running categories of political advertising (see chapter 4.1 above) without fact-checking their content and also resisted calls to prevent micro-targeting connected to it. This divergence in approaches was underlined by a public disagreement that erupted between Twitter, Facebook and the U.S. President Donald Trump in May and June 2020 after Twitter flagged as misleading a tweet from the President about election protocols (Hatmaker, 2020) and hid one of his tweets for 'glorifying violence' (BBC, 2020c). Facebook CEO Mark Zuckerberg explained that Facebook would never take such action against a senior political figure because it was not in the business of being an 'arbiter of truth'. (For a more detailed discussion of this episode, see Chapter 7.1 on normative and ethical responses.)

Filtering happens ex-ante, meaning prior to publication and distribution of content. Restrictions, blocking and removal of the publication and distribution of content can also be ex-post, meaning after content has been initially published. With regards to filtering (prior to publication), platforms make use of hash databases[192] with 'digital fingerprints' of previously flagged content, terrorist content, child sex abuse images, and copyright infringing content to detect and prevent re-uploads. In this context, **YouTube**, **Facebook**, **Microsoft**, and **Twitter** founded the Global Internet Forum to Counter Terrorism

191 https://about.twitter.com/en_us/safety/safety-partners.html

192 Hashing databases refer to the use of a 'hash' or reference to index, retrieve and protect items in a database (Zilavy, 2018).

(GIFCT)[193] to cooperate on technological solutions to combat violent extremism on their platforms.[194]

Internet communications companies also remove, block, or restrict content after receiving machine or human-driven notifications of potentially objectionable material, applying a scale of action depending on the violation at hand. For **WhatsApp**, due to the encrypted nature of the conversations, curbing the spread of disinformation is particularly challenging. WhatsApp started to restrict the number of times a message could be shared to five times. This feature was first introduced in India in July 2018, and subsequently rolled out worldwide in January 2019 (WhatsApp, 2019a). Restrictions on forwarding were tightened further during the COVID-19 crisis, with WhatsApp restricting to once, the number of times that a frequently forwarded message could be re-forwarded (El Khoury, 2020). (See also the discussion below on the 'unintended consequences' of such limitations). It is not evident if the sharing of WhatsApp's metadata with its parent company Facebook has relevance to either side in terms of combatting disinformation.

The Chinese company **WeChat** operates with two service agreements - one for mainland China and another for international users. Longitudinal research from the University of Toronto's Citizen Lab indicates that WeChat monitors content in real-time, removing content on the basis of strings of keywords, URLs and images. They also found that messages of mainland Chinese users are filtered more frequently than those of international users, as is content posted via WeChat's Moments and group chats (as opposed to individual chats) (Ruan et al., 2016; Knockel & Xiong, 2019).

3. Promotion and demotion of content

Another option chosen by Internet communication companies is based on the assumption that "freedom of speech is not freedom of reach" (DiResta, 2018), whereby sources deemed to be trustworthy/authoritative according to certain criteria are promoted via the algorithms, whereas content detected as being disinformational (or hateful or potentially harmful in other ways) can be demoted from feeds. (See Chapter 7.3)

On **Facebook**, clickbait content is tackled by reducing the prominence of content that carries a headline[195] which "withholds information or if it exaggerates information separately" (Babu, Lui & Zang, 2017). Facebook has also committed to reducing the visibility of articles that have been fact-checked by partner organisations and found wanting, and the company adds context by placing fact-checked articles underneath certain occurrences of disinformation.[196] (However, as discussed in chapters 4.1, 7.1, and 5.3, certain categories of political advertising are excluded from these fact-checking efforts). Additionally, the company has begun paying a select group of news outlets for content which is being displayed in a separate 'news' section. At the time of writing, this was still in beta mode and only available to a few hundred thousand U.S.-based users (Kafka, 2020). **YouTube** prioritises content from trusted news organisations in their 'top news' and 'breaking news' shelves as a curatorial act designed to highlight credible content, although this is currently available only to U.S. users (Google & YouTube, 2019).

Snapchat differentiates itself from other social media platforms by "separating social from media" (Spiegel, 2017). The platform offers a separate 'Snapchat Discover' section, which

193 https://www.gifct.org
194 https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en
195 https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/
196 https://www.facebook.com/help/1952307158131536

algorithmically displays stories from select news publishers, content creators and the community, curated and promoted by Snapchat editors (Snapchat, 2017). In 2020, the company removed the U.S. president's feed from its Discover section (Newton, 2020).

4. Disabling and suspension of accounts

In addition to curating content, Internet communication companies tackle what they call inauthentic behaviour and content at an account level. Online disinformation can be easily spread through accounts that have been compromised or set up, often in bulk, for the purpose of manipulation. Several companies prohibit 'coordinated inauthentic behaviour' (including interference from foreign governments) in their terms of service agreements. **Facebook** reports that tackling such behaviour is an ongoing challenge, which they are committed to "continually improve to stay ahead by building better technology, hiring more people and working more closely with law enforcement, security experts and other companies" (Facebook and Instagram, 2019). In this light, the company updated its Coordinated Inauthentic Behaviour (CIB) policy in October 2019, explaining how it acts against "a range of inauthentic activities, whether foreign or domestic, state or non-state" (Gleicher, 2019). Some companies intervene during the registration, as well as in the lifespan of an account. For instance, **WhatsApp** "banned over two million accounts per month for bulk or automated behavior" in a three-month period. Roughly 20% of these accounts were banned at registration (WhatsApp, 2019a). Platform account curation during use tends to follow a graduated approach with warnings before sanctions are imposed. **Line**[197] and many other companies, with the exception of **VK**[198] and **Snapchat**[199], temporarily disable the user's account and only subsequently suspend it, when violation of the terms and conditions of use and/or laws are detected.

Facebook has also been enacting suspensions of group pages that violate its terms of service, both on its site and on Instagram. A recent example is the removal of 790 QAnon Facebook groups, 100 Pages, and 1,500 adverts and the restriction of another 1,950 groups on Instagram (Facebook, 2020). These conspiracy theory sources were deemed to violate Facebook policies because they celebrated violent acts and "had individual followers with patterns of violent behavior". This also included 440 Intagram pages and more than 10,000 Instagram accounts related to QAnon (Facebook, 2020b). The suspension followed from an internal investigation by the company, which showed that membership of these QAnon groups exceeded 3 million (Sen & Zadrozny, 2020).[200]

Suspension of accounts on the grounds of inauthentic behaviour and sharing of disinformation content is not clearcut, as both concepts often overlap in the platform's community guidelines. **YouTube** has the most extensive policy in this regard, which it applies when implementing its rules. If violations to community guidelines are found, content is removed and accounts are given a warning, and up to 'three strikes' within a 90-day period. Warnings are marked on the YouTube channel, but do not have further consequences. 'Strikes' can entail disabling account holders from uploading, creating and editing content on YouTube for one week (1st 'strike'), two weeks (2nd 'strike') and ultimately lead to the removal of the YouTube channel (3rd 'strike'). However, in cases where intervention is required for violations beyond the community guidelines (for

197 https://terms.line.me/line_terms/?lang=ja
198 https://vk.com/licence
199 https://www.snap.com/en-US/community-guidelines/
200 See also this Columbia Journalism Review hosted discussion about the QAnon conspiracy theory and disinformation, featuring a galley of journalsits and researchers https://galley.cjr.org/public/conversations/-MFpKx9fqAg5dUs2DirW

instance in response to "a first-party privacy complaint or a court order"), the strikes policy does not apply and can lead to immediate suspension.[201]

5. Transparency in content moderation and sponsored content

As social media sites and apps are increasingly considered as the de facto online public sphere, it has been argued that content moderation may interfere with an individual's right to freedom of expression. Even though private actors have a right to decide on the moderation policies on their service (within legal boundaries), an individual's right to due process remains. Furthermore, a certain level of insight/transparency should be given to users and third parties into the process of how decisions are made, in order to guarantee that these are taken on fair and/or legal grounds. In 2018, a group of U.S. academics and digital rights advocates concerned with free speech in online content moderation developed the Santa Clara Principles on Transparency and Accountability in Content Moderation.[202] These principles set the bar high for the companies, suggesting detailed standards for transparency reporting, notice and appeal mechanisms. Indeed, as a de facto public sphere, there is a need for dominant entities to use international standards, and not operate more limited ones.

Facebook/**Instagram**,[203] **Google**/**YouTube**,[204] **Twitter**[205], **Snapchat**[206] and **LINE**[207] provide periodic (e.g. quarterly) public transparency reports on their content moderation practices as they align with external (legal) requirements. They tend to be less transparent about their internal processes and practices. All except LINE also run (political) advertising libraries. The libraries of Facebook and Twitter cover all advertisements globally, while Google provides reports for political adverts in the European Union, India and the United States, and Snapchat covers political adverts in the U.S.. It can be noted that Argentina, Canada, the EU, France and India oblige online services (and election candidates) to provide transparency in political advertising. This is a policy response being echoed by many others, including Australia, Belgium, Ireland, Israel, Italy, the Netherlands, New Zealand, OAS, the UK and the U.S. (see Chapter 5.1 Legislative, pre-legislative and policy responses).

As of 22 November 2019, however, Twitter prohibited political advertising globally and issue adverts in the U.S. only. As of April 2020, Reddit[208] has also announced the creation of a U.S.-only political advertising library and emphasised that they forbid deceptive, untrue, or misleading advertising (not only political).

Not all platforms provide transparency on content moderation on their services. As an example, **WeChat** does not provide any notification of filtering. Blocked content remains visible for the sender, but does not appear in the chat of the receiver (Ruan, Knockel, Q. Ng, & Crete-Nishihata, 2016; Knockel & Xiong 2019). There is also a lack of transparency on **VK**. In 2018, Tjournal reported that despite the fact that the VK does not allow advertising of a political nature, the entries of the personal blog of a big city mayor were

201 https://support.google.com/youtube/answer/2802032?hl=en
202 https://santaclaraprinciples.org/
203 https://transparency.facebook.com ; https://www.facebook.com/ads/library/
204 https://transparencyreport.google.com ; https://transparencyreport.google.com/political-ads/home
205 https://transparency.twitter.com/en.html ; https://ads.twitter.com/transparency
206 https://www.snap.com/en-US/privacy/transparency ; https://www.snap.com/en-US/political-ads/
207 https://linecorp.com/en/security/transparency/top
208 https://www.reddit.com/r/announcements/comments/g0s6tn/changes_to_reddits_political_ads_policy/

promoted through the advertising tools of the social network; however, a prominent opposition leader was prevented from posting such content (Likhachev, 2018).

6. User involvement

User involvement requires them to be provided with control over the content, accounts and advertising they see. Internet communication companies offer varying types of involvement, including flagging content for review, prioritising, snoozing/muting and blocking content and accounts, and changing the advertising categories users are placed in. This last tool is only offered by a handful of platforms. **Facebook** allows users to update their 'ad preferences' by changing their areas of interest, as relevant to the advertisers who use this information, and targeting parameters.[209] On **LINE**, users can select their preference for sponsored content on banner adverts on LINE Talk, but not on sponsored content on the LINE timeline or other services (LINE, 2019a; LINE, 2019b). As examples of involvement, **YouTube** offers YouTube Kids and other parental controls to restrict content for children,[210] and **Twitter** allows users to "mute Tweets that contain particular words, phrases, usernames, emojis, or hashtags" to remove them from view on their personalised feeds.[211] Twitter has also been trialling specialised support for what it terms 'frontline defenders' (e.g. journalists trying to combat disinformation on the service and being targeted in the process).

7. Appeal

Finally, in response to curatorial action taken and in line with the Santa Clara Principles on Transparency and Accountability in Content Moderation,[212] it is important from the perspective of protecting freedom of expression that companies have in place procedures to appeal the blocking, demotion or removal of content, disabling or suspension of accounts. This entails a detailed notification of the action, a straightforward option to appeal within the company's own service, and a notification of the appeal decision.

As is evident from the discussion above, responses to disinformation differ. For instance, Facebook reduces the distribution of disinformation rather than removing it, unless it also entails other violations of community standards (e.g. is likely to cause physical harm). At the same time though, as discussed in Chapter 4.1, Facebook exempts from curatorial actions all speech in the form of posts and adverts made by politicians, political parties and affiliates. This hybridity makes it difficult to address the question on appeals connected to disinformation in a direct manner. However, it is clear that, in practice, there is a level of variance in the social media companies' approaches to appeals. Although external appeal to an arbitration or judicial body is theoretically possible in some countries, especially where disinformation intersects with a local legal restriction, few companies offer robust appeal mechanisms that apply across content and accounts, or to notifying the user when action is taken.

In 2018, **Facebook** made changes to its appeals process: previously appeal was only possible for profiles, pages, and groups. As a result, it became possible to appeal in reference to individual posts as well (for nudity / sexual activity, hate speech or graphic

209 https://www.facebook.com/ads/preferences
210 https://support.google.com/youtubekids/answer/6172308?hl=en
211 https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options
212 https://santaclaraprinciples.org/

violence) (Bickert, 2018)[213]. On **WeChat**[214] and **LINE**[215], users are able to request to unblock/unfreeze accounts, but there is no evidence of the possibility to appeal against removal of content. There is no evidence that **Snapchat**[216] or **WhatsApp**[217] have set up appeals processes. This can be particularly problematic from a freedom of expression perspective. For example, one of the known practices deployed by disinformation agents involves false reporting of journalists' profiles and accounts as a means of censorship. (See also the discussion in this chapter and chapter 7.1 on the Facebook Oversight Board).

Efforts by internet communications companies to address disinformation are evolving rapidly but their resistance to responding adequately, on a global scale, and taking publisher-style responsibility for the social and democratic impacts places them at risk of becoming used as factories for 'information disorder' and online abuse (Posetti, 2018b).

6.1.2 Journalistic curatorial interventions

Professional journalism has the discipline of verification at its core.[218] The curation and publication of factual information for mass consumption by news organisations, along with the debunking of falsehoods through accountability journalism (Mayhew, 2020), has been an historically important counter-disinformation function. However, erosion of traditional news gatekeeping functions, along with the 'rise of the audience', and the ubiquity of social media have undermined the power of pre-digital modes of editorial curation as a defence against disinformation (Posetti 2018). *The Guardian's* Editor-In-Chief Katherine Viner has written that "Facebook has become the richest and most powerful publisher in history by replacing editors with algorithms." (Viner, 2017).

Internet communications companies have been described as 'the new gatekeepers' (Bell & Owen, 2017). However, as discussed throughout this report, these companies remain largely reluctant to accept responsibility for traditional news publishing oversight - including verification and curation - despite making decisions to censor some content in a manner that has been criticised as undermining media freedom (Doyle, 2016). Controversies connected to the deletion of information, including historically important news photography, along with suspension of journalists' accounts for sharing news photographs that purportedly breached 'community standards' because they depicted nudity (Kleinman, 2016; Gillespie, 2018). A number of these controversies - which attracted significant media coverage - triggered the processes that ultimately led to the establishment of the Facebook Oversight Board in 2019.

Digital transformation has delivered many benefits, including enhanced opportunities for freedom of expression and access to diverse information. However, it has also fueled unprecedented, ongoing challenges and structural changes to the news industry that favour viral disinformation including by undermining the role of journalistic curation. These include[219]:

213 https://transparency.facebook.com/community-standards-enforcement
214 https://help.wechat.com/
215 https://terms.line.me/line_terms/?lang=ja
216 https://support.snapchat.com/en-US/i-need-help?start=5153567363039232
217 https://www.whatsapp.com/legal/?eea=0#terms-of-service
218 See discussion in chapter 7.1 on normative and ethical responses to disinformation
219 The following examples represent a curation of impacts drawn from: 'News industry transformation: digital technology, social platforms and the spread of misinformation and disinformation' (Posetti 2018), published by UNESCO and available here: https://en.unesco.org/sites/default/files/j._jfnd_handbook_module_3.pdf

- The collapse of the traditional business model for news publishing, leading to mounting outlet closures and mass unemployment within the industry, dramatically reducing curatorial capacity;
- Depletion of newsroom resources (staff and budgets) resulting in less on-the-ground reporting, and affecting fact-checking and editing processes, leading to less scrutiny of information and sources;
- Media convergence: many journalists are now tasked to produce content for multiple platforms concurrently (from mobile to print), further depleting time available for proactive reportage and scrupulous verification;
- Reporters are increasingly required to sub-edit and publish their own content without appropriate review;
- Increased demand to churn out content to feed homepages and social media channels on top of rising deadline pressure, coupled with reduced quality control processes and job losses, exacerbates the weakening of standards;
- Audience expectations of 'on-demand' news, mobile delivery and realtime engagement on social media further increasing pressure on news professionals facing diminishing resources in a never-ending news cycle. Digital-first deadlines are always now, heightening the risk of errors, including the inadvertent sharing of disinformation a or material from spurious sources
- 'Social-first' publishing is commonplace, with reporters curating their own individual newsfeeds on social media accounts to meet audience demand for real-time news. Practices include 'live tweeting', 'Facebook Live' videos, and other journalistic acts which do not necessarily involve editorial oversight (akin to live broadcasting), potentially resulting in a 'publish first, check later' mindset;
- News publishers are struggling to hold onto audiences as barriers to publication are removed, empowering any person or entity to produce and curate content, bypass traditional gatekeepers, and compete for attention – including powerful actors seeking to undermine the credibility of critical reporting;
- Targeted online harassment of journalists (particularly women), their sources and their audiences, distracting and inhibiting them from countering disinformation inside the social media communities where it flourishes;
- Clickbait practices (understood as the use of misleading headlines to entice readers to click on links under false pretences) designed to drive traffic and which have been associated with erosion of trust in professional journalism;
- Pursuit of virality at the expense of quality and accuracy.

The result of all of this is that audiences may not turn to news media in times of crisis and disaster with confidence that they will be served well-curated, reliable, verified information published and shared in the public interest. This has the potential to significantly impede counter-disinformation through institutions specialised in expert editorial curation of content, audiences and information sources. Nevertheless, some media institutions have undertaken effective interventions in this regard.

One example is journalism that reinforces or triggers curatorial responses to disinformation within the social media companies. One such case study is the news outlet Rappler's approach. They built a 'shark tank' database to track disinformation networks online, then reported on their findings, informing internet communications companies of their work. Some of Rappler's forensic digital investigations have contributed to Facebook's actions regarding the takedown of 'coordinated inauthentic posts' as the company describes orchestrated disinformation campaigns (Rappler Research Team, 2018; Posetti et al., 2019a & 2019b; Garside, 2020).

Another example is where fact-checking collaborations between news outlets, internet communications companies, fact-checking organisations and other third party verification experts help to curb disinformation on social media (see detailed discussion of these approaches in chapters 4.1, 4.2 and 5.3 on Monitoring, fact-checking, investigative and electoral responses). These can be considered collaborative responses designed to improve social media curation on the companies' sites/apps. For example, ahead of the national elections in India in April 2019, **WhatsApp** partnered with Proto,[220] a collaborative social enterprise focused on the digital transformation of journalism, on the action-research project 'Checkpoint'. As part of the project, users were invited to report suspected false content to a helpline that would in return generate verification reports. Beyond verifying content, this project was designed to collect data reported from the users that would otherwise have been unavailable due to the encrypted nature of the 'closed' chat app. The data collected was intended to enable analysis of disinformation on the platform circulating virally on WhatsApp, although it is not known if this resulted in WhatsApp banning actors for what Facebook terms Coordinated Inauthentic Behaviour.

6.1.3 What and who do curatorial responses monitor/target?

Firstly, curatorial responses focus on the ***content*** shared on internet communications companies' sites and apps, the material published by journalistic actors, and the ***users/audiences*** of both. However, WhatsApp (owned by Facebook) uses behaviour as a proxy to avoid moderation practices that are content-based, and which would require amending end-to-end encryption policy[221]. Internal to the internet communications companies, machine learning and content moderation staff detect and act on potentially abusive content, often in collaboration with news organisations, while externally, law enforcement, fact checkers and other third parties contribute as well. The flagged content is subsequently verified, deprioritised or removed. In rare cases, prosecutions also ensue as a result.

In terms of targets, curation can signal to users what content is sponsored, as distinct from what is shown through the organic operation of the companies' algorithms. Some measures target a specific category of content and paid-for content. Among the measures analysed in this chapter, several specifically target ***political*** content and political actors, whether in particular electoral periods, or as a general policy. As an example, Facebook/Instagram, Google/YouTube, Twitter and Snapchat showed transparency in how they curated advertising by rolling out libraries of political adverts, but with different

220 https://www.checkpoint.pro.to/

221 End-to-end encryption has an important role to play in upholding freedom of expression and privacy rights in the Digital Age. In fact, the UN Special Rapporteur on the Right to Opinion and Freedom of Expression has identified encryption and anonymity as enablers of human rights: http://www.justsecurity.org/wp-content/uploads/2015/06/Kaye-HRC-Report-Encryption-Anonymity.pdf

(and frequently limited) geographical scope.[222] Regarding the verification of political advertising, the platforms also chose different options. Twitter banned political advertising in November 2019[223], whereas Snapchat claims it approves every political advertisement posted on the platform.[224] Facebook decided not to verify certain categories of political advertising(see chapter 4.1),[225] limiting scrutiny of political disinformation, while Google updated its policy to restrict political micro-targeting (Spencer, 2019).

With regard to other content prone to disinformation, such as ***health*** and public safety, online platforms have also adapted their policies. To curb vaccine misinformation on its services, Twitter decided in May 2019 to redirect users to public health sources when they looked for information on vaccines (Harvey, 2019).[226] More specifically, the World Health Organisation (WHO) has partnered with Internet communication companies to make sure users are provided with authoritative information on the Coronavirus epidemic, while Google and Twitter have worked to ensure that WHO information ranks first in queries.[227] Facebook used information from authoritative sources like WHO and CDC and fact-checkers to limit the spread of verified false information about the virus and committed to restrict hashtags used to spread disinformation about the epidemic on Instagram.[228]

Secondly, curatorial responses target ***accounts*** abusing the terms of service of the companies, and when relevant, where they run up against certain legal provisions. These abusive actors can be individual users, as well as professional communicators and advertisers, perpetrating isolated or coordinated malicious actions. In these cases, accounts are often disabled or suspended.

Third, another option chosen by online platforms is to involve ***users*** in curating some of the content they see. This can be done by giving users the possibility to flag content, block/snooze content and accounts, change settings of algorithmic recommendations, or change advertising categories in which they have been placed. Users can also be offered the possibility to appeal a moderation decision if they consider their content or account has been wrongly blocked, disabled or suspended.

In the case of journalistic actors, they collect and curate content that can help curation by the internet communications companies, as in the examples above, as well as serve their own audiences, which includes those audiences curated as collaborative responders to disinformation on social media sites and apps.

For the latter, access to well-curated accurate information is an important defence against the spread of disinformation. The targets of journalistic curation also include purveyors of disinformation who exploit the comments sections of news publications and their social media accounts, along with those accounts of individual journalists. Curatorial interventions in these cases are limited to pre-moderation curation in the case of news websites' comments, and post-moderation in the case of social media sites like Facebook

222 Facebook/Instagram rolled out a political ads library across the EU in Spring 2019. Similarly, Google/YouTube offers transparency reports on political advertising in the European Union, India and the United States. Following other companies, Snapchat has decided to roll out a political ads library in the US.

223 https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html

224 https://www.snap.com/en-US/ad-policies/political/

225 https://www.facebook.com/business/help/208949576550051?id=288762101909005

226 This policy has been rolled out in the U.S. (in English and Spanish), Canada (in English and French), UK, Brazil, Republic of Korea, Japan, Indonesia, Singapore, and in Spanish-speaking Latin American countries.

227 https://twitter.com/Google_Comms/status/1222991098257108992

228 https://about.fb.com/news/2020/05/coronavirus/

and Instagram. When it comes to Twitter and chat apps, there is no ability to moderate comments, but there is the power to curate followers and limit the amplification of dubious users who tag, retweet and forward content.

There have been noteworthy developments in the area of news organisations' online comment curation, including a trend towards ending comments outright in order especially to minimise disinformation-laced hate speech (WAN-IFRA, 2016).

6.1.4 Who do curatorial responses try to help?

Due to their international presence, the curatorial responses initiated by the internet communications companies are implemented with potentially ***global*** impact. But with growing pressure from regulators and public opinion to react to specific ***local*** contexts (elections, major events, human rights abuses etc.), some measures have been increasingly tailored and implemented locally, sometimes before being rolled out globally. There is also a practice of U.S.-centric responses to online toxicity - with a corresponding neglect of developing countries.

These measures usually apply to ***all users*** of the companies regarding, for example, the flagging of content and content moderation appeal mechanisms (as they are defined in the companies' terms of service, community guidelines and editorial policies). Some measures are more relevant to ***public authorities***, such as flagging suspected illegal behaviour, or suspension of identified accounts, either under legal obligations or the companies' own rules. However, in comparison to other actors, political leaders are often given more hands-off treatment in practice.

Finally, it could be argued that the responses put in place by these companies serve the objective of preserving their activities and ***business models***. To prevent backlash and avoid hard regulation that would likely increase their responsibility for content posted by their users, it is in their own interest to deal internally with issues of disinformation (and hate speech) on their services. It is arguably also in the interests of some to continue accepting misleading political advertising purely from the perspective of profit or strategic interest in having a playing field that advantages disinformation dealers above truth-tellers if this means a hands-off regulatory scenario for the future.

The motivating factors behind curatorial responses differ, depending on whether they result from a voluntary action by the internet communications companies, or from regulatory pressure. Actions undertaken voluntarily by the companies result from the assumption that clear rules and guidelines for users about posting content, ideally together with transparent content moderation rules and empowerment tools, will nudge users towards resisting disinformation content, including that which features elements of hate speech.

Similarly, the companies consider some degree of automated review of content and accounts necessary and appropriate to scale in order to 'clean up' their services without the cost of hiring armies of human curators. To date, automation of moderation processes has mostly been limited to spam, bulk and automated accounts, copyright infringement, and content previously detected as 'abusive' or 'illegal', although lack of transparency in reporting makes this somewhat difficult to assess. This issue is covered in detail in chapter 6.2 which deals with technical/algorithmic responses.

Responses by companies under regulatory pressure are based on the idea that some degree of intervention is necessary to enforce the law, with the final aim to create environments that discourage disinformation tactics, including online abuse. Curation can also help companies avoid legal cases, and works towards fostering and maintaining the trust of the bulk of their users that they are in the hands of 'good corporate citizen' stewards who respect a fiduciary obligation to care for the interests of their customers.

Journalistic actors, on the other hand, are largely motivated in their curatorial responses to comment and social media management by a desire to:

- Live up to their ethical mission and business necessity for verified information;
- Ensure that their audiences have access to accurate and reliable information, while being protected from exposure to damaging disinformation;
- Protect their journalists and websites from attack;
- Protect their audiences from attack through disinformation;
- Ensure the integrity of their journalism.

Additionally, there are technology-based solutions for comment curation, such as those devised by the Coral Project (originally a collaboration between the *Washington Post*, the *New York Times* and Mozilla, now owned by VoxMedia[229]). (See the next chapter - 6.2 - for more on technology's application in counter-disinformation curation).

6.1.5 What outputs do curatorial responses produce?

The outputs resulting from curatorial responses to disinformation vary according to the approach undertaken and the actor/s involved. The number of accounts removed or suspended, comments deleted, content demoted/promoted, filtered or blocked, etc. is sometimes made public by the internet communications companies or news organisations (and other actors publishing public interest information) in the form of transparency pages, blog posts, or selective comments from authorised representatives.

The internet communications companies' transparency reports vary greatly, limiting comparability between them. Similarly, the lack of detail in reporting (such as detailed reasoning for action taken) or even absence of reporting on moderation practices (such as for the (de)prioritisation of content), make it difficult to evaluate the extent and effectiveness of measures taken. When such actions result from self-regulatory commitments overseen by public authorities, they may publish transparency reports, such as in the framework of the EU Code of Practice on Disinformation, and the German Network Enforcement Act (see Chapter 5.1 Legislative, pre-legislative and policy responses). For example, Facebook was fined 2 million Euro by the German Federal Office of Justice in 2019 for lack of transparency in its reporting on the complaints filed and actions taken when tackling hate speech and other criminal offences (Prager, 2019; Zeit, 2019). Finally, the reports can also be drafted by content moderation boards, as Facebook (2019) initially committed to with its Oversight Board.

229 https://coralproject.net/about/

6.1.6 Who are the primary actors in curatorial responses, and who funds them?

The curatorial responses of social media actors are largely implemented by the internet communications companies with their own resources. Reliable figures on platform expenditure on content curation are hard to come by. Although it is an incomplete picture, some detail can be offered on Facebook. For example, The Verge reported that Facebook offers contracts of $200 million for content moderation with external contracting agents (Newton, 2019a). In the U.S., contractors are paid approximately 1/10th of a regular Facebook employee, for work that puts individuals under significant psychological strain, at times resulting in post-traumatic stress disorder. Outsourcing of content moderation to South East Asia, especially the Philippines, is also common among the companies (Dwoskin, Whalen & Cabato, 2019; Newton, 2019b).

Moreover, evidence has emerged that the combination of stress and repeated exposure (Schumaker, 2019) to conspiracy theories and other disinformation are leading to content moderators starting to believe the false content that they are actually meant to be moderating (Newton, 2019c). This firmly places the onus on Facebook and other internet communication companies who rely extensively on content moderators to implement effective steps towards protecting their contractors from the harmful effects of disinformation, as well as towards improving their pay and working conditions.

Further, Facebook has set aside $130 million for the operation of its Oversight Board over the next six years (Harris, 2019). Finally, as part of the Facebook Journalism Project, Facebook also announced that they will launch a $1 million fund to support fact-checking[230] and a $1 million fund to support news reporting[231] on COVID-19. Similarly, Twitter will make $1 million available to protect and support journalists during COVID-19 (Gadde, 2020). Facebook's annual revenue amounted to $70.7 billion in 2019.[232] As the business model of the companies mainly relies on targeted advertising, one could argue that since this advertising scheme is based upon the data collected from users, it is the latter who indirectly finance these costs in the responses to disinformation.

In the case of journalistic actors' curatorial responses to disinformation, these are either funded by the news organisations themselves, by individual journalists acting independently to manage their social media accounts, or via grants from foundations or the internet communications companies that are designed to improve audience curation and community management.

6.1.7 Response Case Study: COVID-19 Disinformation

a. Responses from internet communication companies

There have been unprecedented reactions to the 'disinfodemic' from the internet communications companies to limit the spread of false health-related information and redirect users to authoritative sources (Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b), Measures have included stricter implementation of their policies and the adoption of emergency actions, along with a broadening of application of policies to political

230 https://www.facebook.com/journalismproject/coronavirus-grants-fact-checking
231 https://www.facebook.com/journalismproject/programs/community-network/coronavirus-grants-news-reporting
232 https://www.statista.com/statistics/268604/annual-revenue-of-facebook/

actors in certain cases. The unique situation pushed the companies to work closely together, and even publish a common industry statement endorsed by Facebook, Google, Linkedin, Microsoft, Reddit, Twitter and YouTube, in a move to jointly combat fraud and disinformation on their services.[233]

For the purpose of this case study, we examined the measures taken by these companies. All of them took the initiative to redirect users to reliable information and limit the spread of disinformation. Some of these measures were taken proactively, while others were taken after discussion with public authorities. In the table and text below, additional analysis is provided on a number of the biggest internet communications companies. The XX markings indicate where the online platforms have taken extra measures to curb the spread of COVID-19-related disinformation.

Content / account moderation		Facebook Instagram	WhatsApp	Google YouTube	Twitter
Flagging and review of content	Machine driven	XX		XX	XX
	Human driven	X		X	X
	Third party review	XX	(X)	X	X
	External counsel	X		X	
Filtering, removal, blocking and limiting of content	Re-upload filter	X		X	X
	Restricted content forwarding		X		
	Restrictions based on:				
	• platform rules	X	X	X	XX
	• law enforcement	X	X	X	X
	Proactive removal of disinformation	XX		XX	
Promotion and demotion of content	Promotion of authoritative sources	XX		XX	X
	Demotion of clickbait	X		X	X
Disabling and suspension of accounts	Graduated approach:				
	• warning			X	
	• limited features	X	X	X	X
	• suspension	X	X	X	X

233 https://twitter.com/fbnewsroom/status/1239703497479614466?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1239703497479614466&ref_url=https%3A%2F%2Ftechcrunch.com%2F2020%2F03%2F16%2Ffacebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation%2F

Transparency in sponsored content	Demarcation of sponsored content	X		X	X
	Ad transparency centre	X		X	X
	Removal of adverts capitalising on the crisis situation	XX		XX	
User empowerment	User can flag content for review	X	X	X	X
	User can block/snooze content/ accounts	X	X	X	X
	User can prioritise content/ accounts	X		X	X
	User can change advertising categories s/he is placed in	X		X	X
Appeal	Notice of action	X		X	X
	Possibility to appeal	(X)		(X)	(X)
	Notification of appeal decision	(X)		(X)	(X)

Table 6. *Curatorial responses from internet communication companies to the COVID-19 Disinfodemic*

1. *Flagging and review of content*

In addition to partnerships with fact-checkers, several platforms implemented additional measures to remove flagged content by public health authorities during the pandemic.

To limit the spread of COVID-19, the internet communications companies and government authorities encouraged confinement of workers at home. With a large number of staff working remotely, the companies chose to increasingly rely on algorithms for content moderation. This has been the case for Facebook/Instagram (Jin, 2020), but also Twitter (Gadde & Derella, 2020) and Google/YouTube (Pichai, 2020). As anticipated by the companies, the increase in automated moderation led to many bugs and false positives.[234]

2. *Filtering, removal, blocking and other restrictions of content*

To limit the dissemination of disinformation narratives related to the coronavirus, several of these companies also took a more proactive approach to removing content. Google claimed to proactively remove disinformation from its services, including YouTube and Google Maps. For example, YouTube removed videos that promoted medically unproven cures (Pichai, 2020). Facebook committed to removing "claims related to false cures or prevention methods — like drinking bleach cures the coronavirus — or claims that create confusion about health resources that are available" (Jin, 2020). Also, the company committed to removing hashtags used to spread disinformation on Instagram. Twitter broadened the definition of harms on the platform, to include denial of public health authorities' recommendations, description of treatment known as ineffective, denial of scientific facts about the transmission of the virus, claims that COVID-19 was part of a conspiracy to manipulate people, incitement to actions that could cause widespread panic, or claims that a specific group would be more or never susceptible to COVID-19.

[234] https://twitter.com/guyro/status/1240063821974138881

3. *Promotion and demotion of content and User involvement*

The primary strategies of the internet communications companies to face disinformation related to coronavirus were to redirect users to information from authoritative sources, in particular via search features of the companies' platforms, and to promote authoritative content on homepages, and through dedicated panels. On Facebook and Instagram (Jin, 2020), searches on coronavirus hashtags surfaced educational pop-ups and redirected to information from the World Health Organisation and local health authorities. The WHO and other organisations also granted free advertising credit by several internet communications companies to run informational and educational campaigns. Google also highlighted content from authoritative sources when people searched for information on coronavirus, as well as information panels to add additional context. On YouTube, videos from public health agencies appeared on the homepage (Pichai, 2020). Similarly, when users searched for coronavirus on Tik Tok, they were presented with a WHO information banner (Kelly, 2020a). Twitter, meanwhile, curated a COVID-19 event page displaying the latest information from trusted sources to appear on top of the timeline (Gadde & Derella, 2020). Snapchat has used its "Discovery" function to highlight information from partners (Snapchat, 2020).

4. *Disabling and suspension of accounts*

The companies had not implemented additional measures regarding the disabling and suspension of accounts with regards to COVID-19 disinformation. Nonetheless, Twitter had worked on verifying accounts with email addresses from health institutions to signal reliable information on the topic.[235]

5. *Transparency in content moderation and sponsored content*

The WHO and other authoritative organisations were granted free advertising credit by Facebook and received help for advertising from Google. Regarding sponsored content, most platforms chose to block adverts trying to capitalise on the pandemic. Nevertheless, many scams appeared on social media, leading law enforcement and consumer authorities to warn consumers and call on marketplaces to react quickly.[236]

6. *Appeal*

No specific changes to appeal mechanisms related to COVID-19 have been noted, although the COVID-19 crisis led to workforce depletion and a greater reliance on automated content moderation of coronavirus disinformation. Facebook cautioned that more mistakes were likely and that it could no longer guarantee that users who appealed against automatic removal would have recourse to a human-based review process. Similar announcements were made by Google, Twitter and YouTube. In cases where automation erred (e.g. a user post linking to a legitimate COVID-19 news or websites was removed), the dilution of the right to appeal, and the lack of a robust correction mechanism represented potential harm for the users' freedom of expression rights (Posetti & Bontcheva, 2020a). This weakens one of the key corporate obligations highlighted by the UN Special Rapporteur on the right to Freedom of Opinion and Expression (UN Special Rapporteur on Freedom of Opinion and Expression, 2018b, section IV, pars 44-63).

b. Curatorial responses to the 'disinfodemic' from journalistic actors

Curatorial responses were also a major plank of news organisations' strategies for combatting the 'disinfodemic' (Posetti & Bontcheva, 2020a). Apart from tightening

[235] https://twitter.com/twittersupport/status/1241155701822476288?s=12
[236] https://www.consumer.ftc.gov/features/coronavirus-scams-what-ftc-doing

moderation practices in online comments and heightened awareness about the increased risks on audience engagement on branded social media channels like Facebook, where pre-moderation of comments is not possible, news publishers rolled out specially curated editorial products designed to educate and inform their audiences.

Examples of such journalistic curatorial interventions included:

- Thematic newsletters that curate the best reporting, research and debunking on a scheduled basis[237].
- Podcasts that mythbust through the curation of fact checks, interviews, data reviews, and credible public health information on COVID-19[238].
- Live blogs[239], and regularly updated lists[240] and databases of debunked disinformation from around the world[241].
- Specialised curations that centralise resources, guidelines, and explanatory reporting about doing journalism safely, ethically, and effectively during the pandemic[242].

Additionally, the NGO First Draft compiled a list of how 11 major internet platforms were responding to what they framed as mis-and disinformation around the COVID-19 pandemic[243]. Some major actions identified included deregistering obvious disinformation purveyors, while elevating credible sources through free advertising space and other mechanisms.

As traditional gatekeeper institutions in the production and transmission of content, media institutions face particular challenges related to the 'disinfodemic'. Media diversity is a valuable contribution to society, but some news publishers have been captured by forces that unduly politicise the crisis in ways that approach the level of disinformation. Some journalists are also vulnerable to hoaxes, sensationalism, and the ethically problematic practice of wrongly interpreting a commitment to objectivity through a 'false-balance' approach, where they weigh *untruthful* and *truthful* sources equally and, too often, uncritically (Posetti & Bontcheva, 2020b). These phenomena led to COVID-19 disinformation being legitimised by some news outlets (Moore, 2020; Henderson, 2020). Such system failures work against the role of journalism as a remedy for disinformation, and they reduce the news media's potential to call out wider system failure such as the lack of official information and readiness or the misdirection of public resources.

237 See, for example, the Infodemic Newsletter from CodaStory https://mailchi.mp/codastory/the-infodemic-may-3726181?e=57d6fdb385

238 See, for example, ABC Australia's 'Coronacast' podcast https://www.abc.net.au/radio/programs/coronacast/

239 See, for example, *The Guardian*'s comprehensive liveblogging of the pandemic https://www.theguardian.com/world/live/2020/mar/31/coronavirus-live-news-usa-confirmed-cases-double-china-update-uk-italy-spain-europe-latest-updates

240 See Buzzfeed's living curation of coronavirus myths and hoaxes https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes

241 See the Poynter Institute's curation of factchecks and debunks about COVID-19 https://www.poynter.org/fact-checking/2020/the-coronavirusfacts-global-database-has-doubled-in-a-week-check-out-the-latest-hoaxes-about-covid-19/

242 See the International Center for Journalism's (ICFJ) curated resources to assist reporting on coronavirus https://ijnet.org/en/stories#story:7100

243 https://firstdraftnews.org/latest/how-social-media-platforms-are-responding-to-the-coronavirus-infodemic/

The COVID-19 crisis was also an opportunity for many news publishers and journalists to strengthen their public service through reinforced editorial independence, along with adherence to the highest standards of ethics and professionalism, with strong self-regulatory mechanisms. In this way, journalism was able to demonstrate its accountability to standards, distinguishing itself from the kind of problematic content and interaction prevalent in the expanding space of private and direct messaging (including messaging apps such as WhatsApp), where disinformation and its agents thrive away from the wider public gaze and continue unchecked. News publishers in this mode were able to demonstrate their trustworthiness as a source of facts and fact-based opinion, reinforcing this by exposing organised actors within the 'disinfodemic'. Similarly, they highlighted their important role in ensuring publicly accountable and transparent responses from all actors to both the 'disinfodemic' and the wider COVID-19 crisis.

6.1.8 How are these responses evaluated?

The curatorial responses put in place by internet communication companies primarily consist of self-regulatory measures, and thus do not follow a consistent reporting structure. The guidelines, transparency reports and corporate blog posts or occasional announcements give some rudimentary insight into the decision-making processes of the companies. Evaluation by governments[244], academics (Andreou et al., 2018), media (Lomas, 2020), and civil society groups (Privacy International, 2020) indicates both their value and potential limits of internet companies' curation (see the discussion above on the Santa Clara Principles, and the UN Special Rapporteur in '*Challenges and Opportunities*' below). In some cases, regulators also assess the self-regulatory commitments with a view to potentially developing new regulatory proposals. For example, the EU Code of Practice on Disinformation involves assessment of commitments by the European Commission and regulators, prior to a possible revision or regulatory proposal.

Where legislation obliging online platforms to counter the spread of disinformation has been passed (see Chapter 5.1 Legislative, pre-legislative and policy responses), evaluation criteria can be included more systematically. As an example, in February 2020, the German government approved a regulatory package to update and complement their 2017 Network Enforcement Act (German BMJV, 2020a; German BMJV, 2020b).

In the case of evaluating curatorial responses to disinformation by journalistic actors, there is no systematic process of evaluation, but a variety of industry measures are applicable, spanning the level of individual journalists to peer review processes such as through press councils and professional awards. At the individual journalist and news organisation levels, social media metrics and newsroom analytics measure some outcomes of curation including the reach and 'stickiness' of audience engagement (e.g. time spent with an article, the number of new subscriptions/memberships, follows, shares and comments). This does not necessarily present an accurate impression of impact, because stories or posts with relatively low audience reach may still achieve significant policy impact at the State or intergovernmental level.

Professional awards also recognise the role of editorial interventions in the disinformation crisis. For example, the joint winners of the biggest international award for investigative

[244] See for example, the UK parliament's attempt to scrutinise the internet communications companies' approaches to curating disinformation during the COVID-19 pandemic: https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/sub-committee-on-online-harms-and-disinformation/news/misinformation-covid-19-19-21/

journalism in 2019 (the Global Investigative Journalism Network's Shining Light Award) won on the basis of a series of reports and other curated content that helped expose disinformation networks with links to the state in South Africa and the Philippines (Haffajee, 2019).

6.1.9 Challenges and opportunities

Previous disinformation campaigns have made clear that without curatorial intervention, the services operated by internet communications companies would become very difficult to navigate and use due to floods of spam, abusive and illegal content, and unverified users. As the companies themselves have access to data on their users, they are well placed to monitor and moderate content according to their policies and technologies. Putting strategies in place, such as banning what the companies sometimes refer to as 'coordinated inauthentic behaviour' from their services, or promoting verified content, can help limit the spread of false and misleading content, and associated abusive behaviours. However, policies are best developed through multi-stakeholder processes, and implementation thereof needs to be done consistently and transparently. Monitoring this could also be aided by more access to company data for ethically-compliant researchers.

An approach that favours cooperation and engagement with other stakeholders, including fact-checkers and independent advisory boards, enables external oversight. It also has the potential to keep at bay legal interventions that could unjustifiably curb freedom of expression. This approach aligns with the view of the World Summit on the Information Society, which urges multi-stakeholder engagement in governance issues, ranging from principles through to operational rules. (World Summit Working Group, 2005)

It is difficult to monitor and evaluate the efficacy of curatorial responses in the absence of greater disclosure by the internet communications companies. This has led to growing controversy over the companies identifying, downgrading and deleting content and accounts that publish and distribute disinformation. In parallel, there is concern about special exceptions to these rules made for powerful political figures[245]. For instance, it is not clear how often, or under which circumstances, *ex ante* filtering and blocking of content and accounts takes place on these companies' platforms. Some review and moderation is machine-driven, based on scanning hash databases and patterns of inauthentic behaviour. But it is unclear which safeguards are in place to prevent the over-restricting of content and accounts[246]. This is borne out via controversies connected to inappropriate deletions justified on the grounds of breaching platform rules. Curatorial responses, especially when automated, can lead to many false positives/negatives.[247] Transparency on the frequency and categories of filtering is notably absent, and appeal mechanisms on curatorial responses are generally weak across most of the companies. Taken together, all these raise major concerns from a freedom of expression perspective.

Redress actions can be taken on the basis of existing law and community standards. Yet, a major limitation in the compliance of social media companies with national regulation needs to be noted, as they operate globally and do not necessarily fall into the legal frameworks of the jurisdictions where they operate. Companies prefer to operate at

245 See the earlier discussion in this chapter regarding Twitter, Facebook and the US President, along with analysis of that controversy in chapters 5.3 and 7.1

246 See further details in chapter 6.2 - Technical/algorithmic responses

247 In the context of the coronavirus crisis, Facebook strengthened its moderation on the issue. However, the use of automated anti-spams filters led to the removal of credible sources. https://twitter.com/guyro/status/1240063821974138881

scale in terms of law: they are usually legally based in one jurisdiction, but their users cross jurisdictions. Adherence to national laws is uneven, and in some cases, moderation policies and standards follow the headquarters' interpretation of standards for freedom of expression, more closely than a particular national dispensation. In some cases, this benefits users such as those in jurisdictions with restrictions that fall below international standards of what speech enjoys protection.

At the same time, terms of service, community guidelines and editorial policies often tend to be more restrictive, and thus limit speech, beyond what is legally required at least in the jurisdiction of legal registration (e.g. Facebook's censorship of culturally significant nudity or breastfeeding mothers). Private companies with global reach are thus largely determining, in an uncoordinated manner currently, what is acceptable expression, under their standards' enforcement. This can result in these companies acting as definers, judges and enforcers of freedom of expression on their services. Indeed, any move by these companies in terms of review and moderation, transparency, user involvement and appeal can have tremendous potentially negative implications for freedom of expression.

Complicating this further is that while recognising the role that internet communications companies need to play in curtailing disinformation published on their platforms, there are potential issues with having regulatory power informally delegated by States to these private companies. This is especially the case where this reduces the accountability and judiciability of expression decisions at large that are the responsibility of States, and which should be in line with international human rights standards. This can amount to privatised censorship. Where delegation is explicitly provided by regulations (see chapter 5.1 which deals with legislative, pre-legislative and policy responses), there can be public accountability for these regulations in democracies which respect the rule of law and issues of necessity and proportionality. However, at the same time and to a large extent, for different political, economic and technological reasons, the internet companies are largely left alone to de facto self-regulate content as they see fit.

Freedom of expression concerns that regulated curation could be worse than self-regulated curation in different parts of the world have some validity. However, the self-regulation of curation is still generally legally liable under laws about copyright and child abuse, for example, so the issue is more about the types of regulation rather than regulation per se. Tricky terrain is entered into when regulations criminalise disinformation, particularly when these are vague and/or disproportionate in terms of international human rights standards. However, consumer regulation about data protection and the ability to appeal decisions, as well as regulation for transparency companies report on how decisions are taken, could be less complex from a freedom of expression point of view.

As highlighted in this chapter's introduction, each internet communications company offers different types of services and operates in different ways, which justifies the need for a differentiation in rules regarding the use of their services. Nonetheless, in the absence of harmonised standards and definitions, each company uses its own 'curatorial yardstick', with no consistency in enforcement, transparency or appeal across platforms. Such pluralistic practice may accord with the different platforms and business models, and it can be positive for the exercise of free expression and combatting disinformation, whereas a more centralised and globally enforceable model could risk working against this. In between these two extremes, there is space for the companies to operate their own ethical balance between what they allow to be expressed, and what moderation decisions are made in relation to disinformation and other content that they may deem to be problematic in terms of their policies, and/or is legally fraught in regard to particular jurisdictions.

The Santa Clara Principles[248] point to a possible framework for transparency and accountability in content moderation. The Principles were developed in early 2018 by a group of U.S. academics and digital rights advocates concerned with freedom of expression in online content moderation. They could be self-regulatory but could also contribute to regulatory policy. They suggest standards for transparency reporting, notice and appeal mechanisms. An example of one recommendation they provide on appeals is to ensure "human review by a person or panel of persons that was not involved in the initial decision." The Principles seek to encourage out a high-level human-rights based approach to moderation.

This kind of approach has also been advocated by the **UN Special Rapporteur on the Promotion and the Protection of the Right to Freedom of Opinion and Expression**, who published a Report on a Human Rights Approach to Platform Content Regulation in 2018 (Kaye, 2018). Similar to the UN/OSCE/OAS/ACHPR Special Rapporteurs' Joint Declaration on Freedom of Expression and 'Fake News,' Disinformation and Propaganda (2017)[249], the report points to the need for balancing when restricting freedom of expression (with due regard to legality, necessity and proportionality, and legitimacy), and liability protection for internet communications companies for third party content. The Special Rapporteur raises concerns around content standards. These pertain to vague rules, hate, harassment and abuse, context, real-name requirements, and disinformation. The Report sets the bar high, laying out human rights principles for corporate content moderation (UN Special Rapporteur on Freedom of Opinion and Expression, 2018b, section IV, pars 44-63):

- *Human rights by default, legality, necessity and proportionality, and non-discrimination when dealing with content moderation*;
- *Prevention and mitigation of human rights risks, transparency when responding to government requests*;
- *Due diligence, public input and engagement, rule-making transparency when making rules and developing products*;
- *Automation and human evaluation, notice and appeal, remedy, user autonomy when enforcing rules*; *and decisional transparency*

The Special Rapporteur also raised concern about "the delegation of regulatory functions to private actors that lack basic tools of accountability," indicating that their "current processes may be inconsistent with due process standards, and whose motives are principally economic" (par 17). The report also specified that "blunt forms of action, such as website blocking or specific removals, risk serious interference with freedom of expression" (par 17), and that technological measures that restrict news content "may threaten independent and alternative news sources or satirical content. Government authorities have taken positions that may reflect outsized expectations about technology's power to solve such problems alone" (par 31).

Many of the challenges and opportunities associated with curatorial responses to disinformation from journalistic actors were outlined above in sections 6.1.2 and 6.1.7 of this chapter. They are focused on the erosion of traditional gatekeeper functions

[248] https://santaclaraprinciples.org/; For reflections on freedom of expression safeguards in use of automated content moderation to tackle disinformation online, see Marsden & Meyer (2019). The following paragraphs on the Santa Clara Principles and the UN Special Rapporteur study can also be found in this earlier study provided for the European Parliament.

[249] See also chapters 5.1 and 7.1 if this report for further discussion

in the social media age. Primary among them, are the twin challenges of surfacing and distributing credible, verifiable public interest information amid a tsunami of disinformation, abusive speech, and entertainment-oriented content, along with poor quality and hyper partisan journalism, that together risk drowning out well-crafted and well-curated counter-disinformation content. Curating audiences at scale on open social media channels and in open comments sections - where disinformation, hate speech and abuse flourish - can also be extremely challenging (Posetti et al., 2019b).

Additionally, there are ethical and professional challenges such as misinterpretation of the principle of objectivity, where false equivalency is mistaken as an antidote to bias resulting in the uncritical and equal weighting of *untruthful* and *truthful* sources. The loss of trust associated with system failures in the news media undermine professional journalism's capacity to act as a bulwark against disinformation.

However, these challenges also represent opportunities for news publishers and journalists to mark themselves out as independent, ethical and critical curators of credible, reliable and trustworthy public interest information (Powell, 2020). They also present opportunities to innovate in the area of audience engagement in closed social communities like WhatsApp to help work against disinformation where it circulates in the absence of wider public scrutiny and debunking (Posetti et al., 2019a).

6.1.10 Recommendations for curatorial responses

Given the challenges and opportunities identified above and the considerable freedom of expression implications of curatorial responses, the following policy recommendations can be made:

Individual States could:

- Promote the need for independent multi-stakeholder 'social media councils', similar to press councils in the newspaper sector, along with regulations that require transparency in how internet communications companies interpret and implement their standards, allow for industry-wide complaints and mandate inter-company cooperation to provide remedies (UN Special Rapporteur on Freedom of Opinion and Expression, 2018b, pars 58, 59, 63, 72)[250].

International organisations could:

- Encourage internet communications companies to ensure the curatorial responses that they initiate are appropriately transparent and measurable, support human rights, and are implemented equitably (e.g. avoiding exceptions being granted to powerful political figures) on a truly global scale.

Internet communication companies:

- Could provide detailed and frequent public transparency reports, including specific information on the viewing and spread of disinformation, suspension of accounts spreading disinformation, removals and other steps against disinformation, including demonetisation, as these responses can have significant human rights and freedom of expression implications.

[250] A similar idea is raised in Wardle (2017).

- Establish robust third party/external review mechanisms for content moderation and ensure the ability to appeal decisions, including machine-driven ones. This includes the need to review decisions not to remove content, as well as decisions to delete it.

- Ensure that curatorial responses encourage users to access journalism from independent and professional news organisations or others publishing critical, evidence based public interest information (e.g. independent researchers and bona fide civil society organisations).

- Increase their efforts against orchestrated disinformation-laced attacks on journalists by excluding users who are part of such assaults on press freedom and act as obstacles to efforts to counter disinformation.

- Take steps to ensure appropriate support for content moderators, including training, commensurate wages for work done, and provision for psychological health.

The media sector could:

- Highlight counter-disinformation content (e.g. content that helps educate audiences about the risks of disinformation, helps equip them to resist and counter it where they find it, and gives prominent exposure to important debunks such as COVID-19 mythbusting).

- Experiment with creative means of audience curation and engagement, especially within closed apps where disinformation flourishes.

- Advocate for curatorial disinformation interventions by internet communications companies and relevant governance bodies to take account of international human rights frameworks, and for any restrictions imposed in emergency situations (e.g. COVID-19) to meet the conditions of international standards on the limitation of rights.

- Critically monitor the curatorial efforts of the internet communications companies to aid transparency and accountability.

Note: Further recommendations specific to curating adverts and demonetisation are addressed in Chapter 6.3.

6.2 Technical / algorithmic responses

Authors: Sam Gregory, Kalina Bontcheva, Trisha Meyer and Denis Teyssou

This chapter reviews state-of-the-art algorithms and technology for (semi-) automated detection of online disinformation and their practical utility across the lifecycle of disinformation campaigns including content and source credibility analysis, network spread, measuring impact on citizen beliefs and actions, and debunking methods. To greater or lesser degrees, these technical measures are designed to reinforce or even to implement companies' curatorial or other policy protocols. Use of technical measures outside of the companies, by civil society and/or academics and other actors, is designed to assess issues such as the presence and flow of disinformation (and other kinds of content). This "downstream" character of technical / algorithmic responses means that challenges or opportunities for freedom of expression arising from the technological application may originate in the upstream formal or informal policies at hand. Failure to embed freedom of expression principles at the design stage of a technical response can limit the effectiveness of the response of risk causing unintended negative impacts. At the same time, problems may also arise from a freedom of expression point of view when the technology design logic has little direct connection to policy/purpose logic and operates autonomously of such direction.

These technical / algorithmic responses can be implemented by the social platforms and search engines themselves, but can also be third party tools (e.g. browser plugins) or experimental methods from academic research. Technology discussed in this part of the study includes hash databases, automated ranking, and upload filters, amongst others. The newly emerging technology and knowhow in analysing automatically generated fake content (known as deepfakes or synthetic media) across audio, text, images and video is also reviewed. This chapter also deals with technological means to identify and act on "co-ordinated inauthentic behaviour" and "inauthentic actors", an approach which is different from and complementary to content identification. It consists of technological identification of patterns that tend to correlate with disinformation campaigns.

Additionally the strengths, weaknesses and gaps in a range of other content verification and media forensics approaches are analysed. One particularly important challenge is how to balance interests in algorithm transparency (for example, to ensure that algorithmic choices are verifiable, and implicit and explicit biases understood), against the danger of weakening algorithm effectiveness, which would allow disinformation actors to exploit weaknesses and devise evasion strategies. Another issue is accessibility to tools dependent on algorithmic approaches.

6.2.1 Who and what are the targets of technical and algorithmic responses?

Technical and algorithmic responses monitor the scope and nature of disinformation, utilising automation as a support to decision-making within internet companies and for third parties. They provide approaches to assess the credibility of content items and sources, and the media integrity of new forms of synthesised media, as well as monitor flow of information and computational activity such as use of bots.

6.2.2 Who do technical and algorithmic responses try to help?

Technical responses primarily support several stakeholders: Internet communications companies, as well as media, fact-checkers and investigators. Tools for image-sharing, video-sharing, search and messaging platforms enable the Internet companies themselves to conduct semi-automated processes of detecting messages, agents and how contents spread, as well as provide information to other parties (e.g. third party fact-checkers). A related set of tools supports the processes of journalists, media, fact-checkers and investigators engaging in specific investigations or documenting scope of disinformation on platforms.

Most automated tools in the disinformation detection space are currently suited to provide input to human decision-making - either at a content item level or assessing a pattern of actor behaviour. At the content level, they provide information to enable human analysis of provenance and manipulation. At the actor level, they provide information on potential bot or troll activity and suspicious networked activity.

The assumption behind technical and algorithmic approaches is that they can reduce the presence and sharing of disinformation and the incentives for disinformation actors. Their current theory of change is that given a massive volume of information and the need to both detect coordinated campaigns or individual manipulations that are not easily discernible by humans, automated tools can assist in both triaging decision-making, reducing duplicative attention and speeding up individual decisions and provision of information. However, in the longer-term, it seems likely that an aspiration is to develop more effective algorithmic and machine learning-driven approaches that reduce the need (and personnel and financial resources required) for human moderation and analysis, and thus allow for more automated curation of content without reference to human moderators as is the case with existing approaches to much online violent extremism.

The move to more automated content moderation forced by COVID-19 and the need to work with a reduced and remote human workforce as Facebook[251], Twitter[252] and YouTube[253] have stated, will likely provide insights in the short-term (provided the companies offer some transparency on what occurs in this forced experiment). In their blog on this issue Facebook notes that with "a reduced and remote workforce, we will now rely more on our automated systems to detect and remove violating content and disable accounts. As a result, we expect to make more mistakes, and reviews will take longer than normal, but we will continue to monitor how our systems are performing and make adjustments." This reflects an understanding that currently automated systems are not a replacement for human oversight, and require robust corrections and appeals (as has been highlighted by the UN Special Rapporteur on promotion and protection of the right to freedom of opinion and expression (Kaye, 2018)).

6.2.3 What output do technical and algorithmic responses publish?

In general, unlike automated systems built to detect child exploitation imagery or violent extremist content which remove content largely without human oversight over each

251 https://about.fb.com/news/2020/03/coronavirus/#content-review
252 https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html
253 https://youtube-creators.googleblog.com/2020/03/protecting-our-extended-workforce-and.html

decision, systems for detecting disinformation at scale provide information for subsequent human processes of decision-making within internet companies on responding to disinformation campaigns or making labelling, downranking or removal decisions on specific content or accounts.

Although most major internet companies now produce transparency reports on levels of content or account takedowns as well as investigatory reports on outcomes in countering particular disinformation campaigns (see section 4.2 for further detail) these reports do not include in-depth transparency on the implications of their use of algorithms, machine learning and other forms of automated decision-making in regard to human rights. Nor do they explain on what criteria these methods are considered effective interventions. The extent of disclosure typically includes broad figures for usage and implementation of automated systems - for example in a recent report[254] Facebook notes its ability to identify 99% of fake accounts proactively (i.e. automatically without human reporting). Platforms argue that this is the appropriate level of transparency given the adversarial nature of content moderation and how 'bad actors' will try and exploit an understanding of the algorithms they use for moderation.

A study by Ranking Digital Rights looked into the issue of transparency in relation to recommendation engines (Ranking Digital Rights, 2020). It reviewed five internet companies including Apple (iOS), Google (Search, YouTube, Android), Facebook (Facebook), Microsoft (Bing, OneDrive) and Twitter, and found governance gaps and weak human rights due diligence. The report notes that "none of the five U.S.-based platforms evaluated make explicit public commitments to protect human rights as they develop and use algorithmic systems" and that "companies operating major global platforms do not provide evidence that they are conducting risk assessments that enable them to understand and mitigate human rights harms associated with how their use of algorithmic systems and targeted advertising-based business models affect internet users". Only one U.S. company (Microsoft) disclosed that it conducts impact assessments on its development and use of algorithmic systems. None of the eight companies in the study disclosed whether they conduct risk assessments on how their targeted advertising policies and practices affect users' freedom of expression and information rights, or their right to privacy or to non-discrimination.

Third-party systems to complement content verification or identify new forms of synthesised media vary in the degree of sophistication of their outputs. A number of third-party tools such as INVID and Assembler integrate a range of open-source tools into dashboards to assist professional journalists and investigators.

6.2.3.1. Intra-company approaches on social media, video-sharing, search engines and messaging for (semi-)automated detection of online disinformation campaigns, including automated tools for detection, hash databases and upload filters

Internet companies deploy a range of automated detection models for content types on their services. These include tools for tracking the organic and artificial spread of information as well as for identifying content that meets criteria for down-ranking, labelling or removal.

[254] https://transparency.facebook.com/community-standards-enforcement

Automated tools for detecting and managing disinformation behaviour

Automated content recognition can be used either to make and implement automated judgements or to assist humans in making decisions on content moderation or identification of patterns. As noted in the EU 'Regulating Disinformation with Artificial Intelligence' report (Marsden & Meyer, 2019), "within machine learning techniques that are advancing towards AI, automated content recognition (ACR) technologies are textual and audio-visual analysis programmes that are algorithmically trained to identify potential 'bot' accounts and unusual potential disinformation material." The report recognises that moderating content at larger scale requires ACR as a supplement to human moderation (editing), but states that using ACR to detect disinformation is prone to false negatives/positives due to the difficulty of parsing multiple, complex, and possibly conflicting meanings emerging from text. If inadequate for natural language processing and even for audiovisual material including 'deep fakes' (fraudulent representation of individuals in video), ACR does have more reported success in identifying 'bot' accounts, according to the report.

Although the actual detection algorithms utilised within platforms for detecting inauthentic content or behaviour are not available for public scrutiny Twitter has integrated detection approaches for whether an account uses a stock or stolen avatar photo, stolen or copied profile text, or misleading profile location (Harvey & Roth, 2018). Facebook has fewer automated bot accounts but needs to identify more sock puppets (multiple false accounts with a real human behind them) and impersonation accounts instead. Identifying these automatically is much harder than finding bots (and sometimes impossible), due to the more authentic human-driven behaviour (Weedon et al., 2017). State-of-the-art research on bot detection methods uses predominantly social behaviour features - such as tweet frequency, hashtag use, and following a large number of accounts while being followed by just a few (Varol et al., 2017; Woolley & Howard, 2016; Cresci et al., 2016). There are also approaches that detect bots based on the high correlations in activities between them (Chavoshi et al., 2017).

Wikipedia, which is built on user-generated knowledge contributions, uses bots (i.e. automated agents)[255] to 'patrol' its pages and identify behaviour deemed to be *deliberately "intended to obstruct or defeat the project's purpose*, which is to create a free encyclopedia, in a variety of languages, presenting the sum of all human knowledge"[256]). The Wikipedia community has made a series of proposals on how to create bots to deal with sock puppet accounts used to perform edits, as might occur in the context of a coordinated disinformation campaign), however these do not appear to have been implemented.

Automated tools for content identification and removal including hash databases and upload filters

Automated tools for content removal such as hash databases and fingerprinting are primarily used in the context of child exploitation imagery, copyrighted images (e.g. YouTube Content ID) and violent extremist content, particularly in the context of legal mandates to identify and remove this content. A hash database enables platforms to identify duplicates or near duplicates, based on matches to existing content items in a database.

[255] https://en.wikipedia.org/wiki/Wikipedia:Bots
[256] https://en.wikipedia.org/wiki/Wikipedia:Vandalism

Hashing is a technique that involves applying a mathematical algorithm to produce a unique value that represents any set of bits, such as a photo or video. There are a variety of hashing approaches including hashing every frame of a video or regular intervals of frames, or hashing subsections of an image. These hashing techniques can help detect manipulation, such as whether an image was cropped, and help identify and verify subsets of edited footage. Tools such as PhotoDNA technology used across companies for child exploitation imagery calculate hash values based on the visual content of an image (by converting the image to black and white, resizing it, breaking it into a grid, and looking at intensity gradients or edges) and so are better at detecting media with alterations and edits, not just exact copies.

Until recently there has been no official coordinated mechanism between Internet companies for monitoring disinformation or for utilising a shared hash or fingerprinting approach in this area, unlike in the case of violent extremism where coordination takes place through entities such as the Global Internet Forum to Counter Terrorism (GIFCT) where most major companies are represented. In March 2020, Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter and YouTube jointly announced that they were working closely together on COVID-19 response efforts and "jointly combating fraud and misinformation about the virus". It is not clear whether this includes a shared hash approach (Facebook, 2020a). It is also not clear how such an approach, if broadened beyond misinformation and disinformation around coronavirus, might bridge the differing policies/community standards of companies in this area (for example, in how they handle political adverts containing falsehoods, or how they manage manipulated media) or the range of ways in which mis/disinformation content shifts as users edit and change it. Similarly the coordination under the Trusted News Initiative between major media and platform internet companies does not appear to include a hashing or matching approach.

Upload filters are often used in combination with hashing and fingerprinting. These assess content at point-of-upload to prevent sharing, and are less utilised in the context of disinformation. There are significant freedom of expression concerns around utilisation of hashing and fingerprinting approaches, particularly in combination with upload filters. These concerns include transparency around how any given image is added to a hash or fingerprint database, as well as concerns around how context is considered around an image (as with genuine content distributed in ways that perpetuate disinformation, for example with an inaccurate social media comment or description). As two researchers note, "Automated technologies are limited in their accuracy, especially for expression where cultural or contextual cues are necessary. The illegality of terrorist or child abuse content is far easier to determine than the boundaries of political speech or originality of derivative (copyrighted) works. We should not push this difficult judgement exercise in disinformation onto online intermediaries" (Marsden & Meyer, 2019).

Concerns around upload filters (and a reason why they are not currently fit for usage in disinformation monitoring) reflect the fact that upload monitoring software cannot distinguish intent such as satire and parody that may repurpose existing content (for further examples see Reda, 2017). Compounding the concerns is the lack of transparency on what content is caught in these filters. To-date upload filters are being used in other areas of content moderation - particularly within copyright enforcement as well as in an increasing manner in the counter-terrorism and violent extremism area - but not in disinformation.

Tools for media and civil society to engage with platforms' systems

Some internet companies also invest in tools to enable third-parties to better contribute to identification or fact-checking of content. As discussed in Chapter 4.1, Facebook

supports a network of third-party fact-checkers who are provided with a queue of stories, both flagged by users and as identified by Facebook internal content review teams. In addition, fact-checkers have the option of adding ones they themselves identify to check for credibility (although it is not automatic they will be paid for this work). Facebook says that it then reduces by 80% the visibility of stories deemed to be false by the fact-checkers (DCMS HC 363, 2018b) as well as reduces the reach of groups that repeatedly share misinformation (Rosen & Lyons, 2019).

Claim Review[257] is a web page markup schema developed by Google and the Duke Reporters' Lab to enable easier tagging of stories with relevant information on the underlying fact that has been checked, who said it and a ruling on its accuracy. A version of this approach - MediaReview - is now being developed to enable fact-checkers to better tag false video and images (Benton, 2020).

As discussed in Section 7.3 Empowerment and Credibility Labelling Responses, a range of companies are considering the possibility of content authentication, attribution and provenance tracking tools on their properties, and the development of authenticity architecture. An example would be the Adobe, Twitter and New York Times Content Authenticity Initiative, which has a goal to create an open and extensible "attribution framework ... that any company may implement it within their respective products and services" (Adobe, 2019).

6.2.3.2. Tools for media and civil society understanding disinformation agents, intermediaries and targets, and enhancing processes for evaluating manipulation and fact-checking

Third-party detection of disinformation agents, behaviour and networks

A key aspect of disinformation analysis is analysing the originating agents of the disinformation campaigns, the other key agents involved, and the explicit or implicit network connections between them. An essential aspect of that is the trustworthiness and credibility of these disinformation agents. Some researchers refer to this as "source checking", and argue that it is hugely important, while currently overlooked, especially in terms of assistance from automated tools and approaches (Wardle & Derakhshan, 2017). Journalism research has proposed several metrics for assessing the quality of news and online media, such as partisan bias, structural bias, topical bias, and source transparency (Lacy & Rosenteil, 2015). However, there are currently no automated methods for calculating these. Automated identification of media bias in news articles has received attention in a recent survey (Hamborg, Donnay & Gipp, 2018). Such content-based source trustworthiness indicators complement the currently better understood indicators from bot detection research. A number of these initiatives built on assessing credibility of actors, e.g. the Global Disinformation Index[258], are discussed in other sections (in particular, Section 7.3).

Disinformation agents are often not acting independently, even though this could be hard to establish sometimes. In order to give the impression that a large number of independent sources are reporting in different ways on the same 'facts', some disinformation sites and/or sock puppet accounts reuse and republish other sites' content, in a practice known as information laundering (Starbird, 2017). Journalists currently lack easy-to-use tools that show which alternative media sites or social network accounts have reused content from another. This is important, since hyper-partisan media and sock

257 https://schema.org/ClaimReview
258 https://disinformationindex.org/

puppets are repackaging and/or republishing content in an attempt to acquire credibility and gain acceptance through familiarity. So far, research has focused primarily on studying retweet and mention patterns between such false amplifiers, e.g. in the 2016 U.S. Presidential Election (Faris et al., 2017), but technology for much more in-depth analysis is needed.

Third party automated message/content analysis

Start-ups working on detection approaches drawing on AI to assess either content quality or indicators that a content item is fabricated include Factmata[259] and Adverifai[260]. Additionally, coalitions like the Credibility Coalition have identified content-based indicators for message credibility as a starting point for potential extensions to existing web schema standards. Key disinformation-related content indicators include clickbait titles and some logical fallacies. These approaches overlap with questions discussed in section 7.3 and are as yet not automatically generated.

Third-party tools for detection of bots, computational amplification and fake accounts or to create aggregated or machine-learned based content trust information

While the major Internet companies remain opaque in terms of their processes for bot detection, there are a number of tools developed by companies, civil society and academia.

A widely used Twitter bot detection service is Botometer[261] (previously BotOrNot), which is provided free of charge by Indiana University. Users can check the bot likelihood score of a given Twitter account, based on its user profile information, friends, and followers. Usage is subject to Twitter authentication and rate limiting on how many requests can be made of the Twitter API. In general, as research-based methods can only use the publicly disclosed data about Twitter accounts, there are concerns regarding how accurate they can be, given that human curators can often struggle to identify bots from public Twitter profiles alone, and do make errors of misattribution. This is set to become even harder, as more sophisticated bots are starting to emerge. Recent work reviews the challenges of automated bot detectors over time, noting problems of variance in terms of false positives and negatives, particularly outside of English language resulting in studies that "unknowingly count a high number of human users as bots and vice versa" (Rauchfleisch & Kaiser, 2020).

In Brazil during elections, a team of researchers at UFMG implemented a 'Bot o Humano' using access to Twitter's API to provide a detection service[262] focused on how bots drive trending topics. The researchers also provided related services to monitor public political WhatsApp groups[263] (Melo & Messias et al., 2019) subsequently also available for use in India and Indonesia) and to monitor Facebook Ads (Silva & Oliveira et al., 2020). Commercial providers also provide services in this space, including WhiteOps.[264]

Third-party research and tool development has primarily focused on Twitter bots, due to Facebook API restrictions. The key enabler for these projects is data on proven bots and sock puppets (falsified online identities that are operated by humans) and all their social media data (e.g. posts, social profile, shares and likes). As with all machine learning

259 https://factmata.com/
260 https://adverifai.com/
261 https://botometer.iuni.iu.edu/#!/
262 http://www.bot-ou-humano.dcc.ufmg.br/
263 http://www.monitor-de-whatsapp.dcc.ufmg.br/
264 https://www.whiteops.com/

processes, this data is necessary for the training of the algorithms for bot and sock puppet detection. Many of these datasets were created by academics (e.g. the DARPA Twitter Bot Challenge (Subrahmanian et al., 2016) and the Bot Repository[265]). To-date, only Twitter has publicly released significant datasets[266] to help independent researchers in this area.

Existing methods from academic research are yet to reach very high accuracy, as they often operate only for publicly accessible account data (e.g. account description, profile photo). This may change with the early 2020 release by Facebook and Social Science One of an extensive dataset of URLs[267] shared on Facebook, including data on interaction and if these posts were flagged for hate speech or fact-checking. The social media companies often make use of additional account-related information, including IP addresses, sign-in details, email accounts, and browser caches, which all make the task somewhat easier. As Twitter describes their own proprietary process, "we work with thousands of signals and behaviors to inform our analysis and investigation. Furthermore, none of our preemptive work to challenge accounts for platform manipulation (up to 8-10 million accounts per week) are visible in the small sample available in our public API" (Roth, 2019).

A number of commercial entities provide related network analysis tools (e.g. Graphika[268]), while upcoming government funded initiatives in the U.S. such as SEMAFOR focus on multi-modal identification of disinformation using physical, semantic, visual and digital integrity indicators.

Tools to assist 3rd-party fact-checking

A number of automated fact-checking tools are being developed by fact-checking organisations and start-up companies, e.g. FullFact[269], Duke University's Reporters Lab[270], Factmata[271], Chequado[272], ContentCheck[273]. The aim is to assist the human fact-checkers in tasks, such as automatic detection of factual claims made by politicians and other prominent figures in TV transcripts and online news, e.g. Full Fact's Live tool[274] and Duke's Tech&Check,[275] which uses Claimbuster (Funke, 2018).

Other automation tools offer tracking mentions of already known false claims, e.g. Full Fact's Trend tool, and automatic checking of simple numeric claims against authoritative databases, e.g. Full Fact Live.

Complementary to these are database and crowd-sourced efforts to generate databases of either sources of disinformation or existing false claims and fact-checks. These include efforts like Storyzy[276] which has a database of fake news sites and video channels (30,000 disinformation sources, by early 2020), and WeVerify,[277] which is building a blockchain database of known false claims and fake content, as well as sites like Rbutr[278] that, rather

265 https://botometer.iuni.iu.edu/bot-repository/
266 https://about.twitter.com/en_us/advocacy/elections-integrity.html#data
267 https://dataverse.harvard.edu/dataverse/socialscienceone
268 https://www.graphika.com/
269 https://fullfact.org/
270 https://reporterslab.org/
271 https://factmata.com/
272 https://chequeado.com/
273 https://team.inria.fr/cedar/contentcheck/
274 https://fullfact.org/automated
275 https://reporterslab.org/tech-and-check/
276 https://storyzy.com/about
277 https://weverify.eu/
278 http://rbutr.com/

than fact-check, provide community-generated links to rebuttal pages. Automated fact-checking tools, e.g. Full Fact Live[279] and Duke's Tech&Check[280], also check incoming claims against existing fact-checks stored either in internal databases and/or assembled automatically based on trustworthy, publicly shared fact-checked claims tagged with the open Claim Review standard schema.

Automated fact-checking methods based on Natural Language Processing (NLP) and AI-based techniques are also being researched. One of the seminal approaches focused on identifying simple statistical claims (e.g. the population of the UK is 60 million people) and checking their validity against a structured database (Vlachos & Riedel, 2015). While the accuracy of these methods is improving continuously, thanks to the creation of large datasets of validity-annotated textual claims (Thorne, Vlachos et al., 2018), they are still considered insufficient for practical use (Babakar & Moy, 2016). However, as more and more human-verified claims are shared openly in machine-readable formats, e.g. Claim Review, these will help NLP and AI fact checking algorithms reach maturity. For the time being, as noted by a Reuters Institute report on automated fact-checking (AFC): "Both researchers and practitioners agree that the real promise of AFC technologies for now lies in tools to assist fact-checkers to identify and investigate claims, and to deliver their conclusions as effectively as possible" (Graves, 2018).

Semi-automated tools to complement content verification

Content verification is concerned with verifying whether an image, video, or a meme has been tampered with or promotes false information. Some of the best known tools have focused on crowdsourced verification (e.g. CheckDesk, Veri.ly), citizen journalism (e.g. Citizen Desk), or repositories of checked facts/rumours (e.g. Emergent, FactCheck). Currently, the most successful verification platforms and products include SAM[281], Citizen Desk[282], Check[283], and Truly Media[284]. There are also some browser tools and plugins aimed at journalists, e.g., the InVID/WeVerify plugin[285] and Frame by Frame[286] (video verification plugins), Video Vault[287] (video archiving and reverse image search), RevEye[288] (reverse image search), Jeffrey's Image Metadata Viewer[289] (image verification), NewsCheck[290] (verification checklist). Plugins offering web content and social media monitoring include Storyful's Multisearch[291] plug-in for searching Twitter, Vine, YouTube, Tumblr, Instagram and Spokeo, with results shown in separate tabs, without cross-media or social network analysis; and Distill[292], which monitors web pages.

279 https://fullfact.org/automated
280 https://reporterslab.org/tech-and-check/
281 https://www.samdesk.io/
282 https://www.superdesk.org/
283 https://meedan.com/en/check/
284 https://www.truly.media/
285 https://weverify.eu/verification-plugin/
286 https://chrome.google.com/webstore/detail/frame-by-frame-for-youtub/elkadbdicdciddfkdpmaolomehalghio
287 https://www.bravenewtech.org/
288 https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf
289 http://exif.regex.info/exif.cgi
290 https://firstdraftnews.org/latest/launching-new-chrome-extension-newscheck/
291 https://chrome.google.com/webstore/detail/storyful-multisearch/hkglibabhninbjmaccpajiakojeacnaf
292 https://chrome.google.com/webstore/detail/distill-web-monitor/inlikjemeeknofckkjolnjbpehgadgge

With respect to photo, image, and video forensics, there are a range of tools e.g. Forensically[293], FotoForensics[294], the Image Verification Assistant[295] developed in the REVEAL FP7 EU project, and the InVID/WeVerify video and image verification plugin[296] (further discussed below). The functionalities currently being offered are based on algorithms that highlight tampered areas, metadata categorisation and analysis, and near-duplicate retrieval based on keyframe matching through reverse image search (typically through Google). All of these tools are limited, particularly when it comes to reviewing media that is of lower resolution, and/or has been compressed or shared via one or more social media/video-sharing platforms. Additionally, forensic attribution typically requires a significant level of technical skill.

The European Union has funded, through its Horizon 2020 framework 5, three year long "innovation actions" and a coordination and support action tackling specifically disinformation. These initiatives include the following:

The EUNOMIA project[297] aims to create a social media companion in both mobile and desktop versions, to assist users in determining which social media user is the original source of a piece of information, how this information spreads and is modified in an information cascade, and how likely the information is trustworthy. EUNOMIA's technologies will be tested in specifically created new instances of the Mastodon micro-blogging platform and Diaspora social network with users participating for the experimental evaluation. The EUNOMIA consortium has 10 partners from 9 EU countries.

The Provenance project[298] wants to enable citizens to evaluate online content while developing digital literacy competencies. At the same time, Provenance plans to help content creators to secure their original work from misuse and manipulation, by registering the original work in a blockchain ledger, tracking how it spreads, and identifying any manipulations that occur later on. The Provenance consortium gathers six partners from four EU countries.

The Social Truth project[299] focuses on creating an open and distributed ecosystem and content verification services to check sources of information during the production process, to provide a digital companion (a chat bot) to help with content verification, as well as search engine rankings and advertising preventions for fraudulent sites. To detect disinformation, Social Truth uses both AI technology and content verification trust and integrity based on blockchain technology. The Social Truth consortium brings together 11 partners from six EU countries.

WeVerify[300] (already mentioned above) aims to develop intelligent human-in-the-loop content verification and disinformation analysis methods and tools. Social media and web content will be analysed and contextualised within the broader online ecosystem, in order to expose fabricated content, through cross-modal content verification, social network analysis, micro-targeted debunking, deep fakes detector and a blockchain-based public database of known fakes. WeVerify tools are integrated in Truly Media (a commercial verification tool) and in the InVID/WeVerify verification plugin, an open-source verification

293 https://29a.ch/photo-forensics/#forensic-magnifier
294 http://fotoforensics.com/
295 http://reveal-mklab.iti.gr/reveal/
296 https://weverify.eu/verification-plugin/
297 https://www.eunomia.social/
298 https://www.provenanceh2020.eu/
299 http://www.socialtruth.eu/
300 https://weverify.eu/

toolbox widely used by the fact-checking community. WeVerify gathers seven partners from six EU countries.

SOMA[301] is a coordination and support action (CSA) that established a Social Observatory for Disinformation and Social Media Analysis to support researchers, journalists and fact-checkers in their fight against disinformation. At the core of the SOMA Disinformation Observatory is a web-based collaborative platform (Truly.media) for the verification of digital (user-generated) content and the analysis of its prevalence in the social debate. A linked DisInfoNet Toolbox aims to support users in understanding the dynamics of (fake) news dissemination in social media and tracking down the origin and the broadcasters of false information. SOMA gathers five partners from three countries.

The Fandango project[302] started one year before the previous projects and runs until the end of 2020. It aims at automating disinformation detection and fact-checking through big data analysis, linguistic and network approaches. Fandango plans to build a source credibility scores and profiles module, a misleading messages detection module, a fakeness detector, copy-move detection tools for image and video analysis and a social graph analysis module. FANDANGO gathers eight partners from five countries.

The U.S. government via its DARPA MediFor[303] Program (as well as via media forensics challenges from NIST[304]) continues to invest in a range of manual and automatic forensics approaches. These include refinements on existing approaches based on discrepancies in the JPEG/MPEG for identifying when other elements have been copy-pasted within an image or whether an element has been spliced from another image file. They also include tracking camera identifiers based on the PRNU (a measure of the responsiveness to light of each cell in the sensor array of a camera that provides a unique 'fingerprint' of a camera when taking an image). Some of these approaches overlap with the provenance approaches described in chapter 7.3 – for example, the eWitness tool for provenance tracking leaves designed forensic traces as part of its technology (Newman, 2019a), while some of the controlled capture start-ups use computer vision (scientific techniques related to image identification and classification) to check for evidence of re-capture of an existing image.

Most of the algorithms under development in programs like the DARPA Medifor program and other related media forensics funding programs have not yet been made available as user-facing tools. Alphabet's Jigsaw subsidiary released Assembler, an alpha tool[305], to selected journalists in early 2020 that provides tools for conventional media forensics, as well as for detecting synthetic faces generated with a tool known as StyleGAN.

Some of the most accurate tools tend to combine metadata, social interactions, visual cues, the profile of the source (i.e. originating agent), and other contextual information surrounding an image or video, to assist users with the content verification task. These semantic approaches align most closely with how OSINT and visual verification practices are carried out by journalists and investigators. Two of the most widely used such tools are the InVID/WeVerify plugin[306] (Teyssou et al., 2017) and the Amnesty International

301 https://www.disinfobservatory.org/
302 https://fandango-project.eu/
303 https://www.darpa.mil/program/media-forensics
304 https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018
305 https://jigsaw.google.com/assembler/
306 https://weverify.eu/verification-plugin/

Youtube Data Viewer[307]. The YouTube Data Viewer extracts metadata listings and offers image-based similarity search using keyframes.

Tools for detection of new forms of algorithmically-generated manipulated media.

To date there are no commercially available tools for detecting a wide variety of new forms of AI-manipulated audiovisual media known as deepfakes and/or 'synthetic media', nor have the platforms disclosed the nature of the tools they are deploying.

Several approaches are being developed however, including a number that rely on either further developments in media forensics, or in utilising the same forms of neural networks that are frequently used to generate deepfakes but here within the detection process. Other forms of detection utilise machine learning but draw on techniques of questioning and interrogating the semantic integrity of images and stories to identify manipulation (Verdoliva, 2020).

Detection approaches to the new generative adversarial network (GAN)-based creation techniques that are used to create deepfakes and other synthetic media can utilise the same technical approach to identify fakes (Gregory, 2019). In early 2020, the first tools were released as part of the Jigsaw Assembler noted above and we should anticipate that some will soon enter the market for journalists either as plug-ins or as tools on platforms in 2020. These tools will generally rely on having training data (examples) of the forgery approach, so they will not necessarily be effective on the very latest forgery methods. As an example, forensics projects such as FaceForensics++ generate fakes using tools like FakeApp and then utilise these large volumes of fake images as training data for neural nets that do fake-detection (Rössler et al., 2018). Major companies have however begun to invest also in supporting independent research as well as the generation of datasets to facilitate solution developments. Examples in this context include Google's work with the Face Forensics project (Dufour & Gully, 2019), and on synthesised audio (Stanton, 2019), as well as the Deepfakes Detection Challenge (Schroepfer, 2019) launched by Facebook, Microsoft, Amazon, the Partnership on AI and a range of academics.

Other approaches in this area also look at evolutions in media forensics to identify the characteristic image signatures of GAN-generated media (Marra et al., 2018) (similar to the PRNU 'fingerprints' of conventional cameras). Outside of programmes like the DARPA MediFor partnership, a number of commercial companies and academic institutions are working in the area of GAN-based detection including (and not limited to) DeepTrace Labs[308], Faculty AI, [309], WeVerify[310] and Rochester Institute of Technology[311]. Key questions around these tools include how well they will work for different types of manipulation, how robust they will be as the forgery processes evolve and improve and how they will present their results in interpretable and useful ways to journalists and users. The recent report of the Partnership on AI's Steering Committee on Media Integrity, which provided oversight on the Deepfakes Detection Challenge, provides further guidance on how to operationalise these concerns in developing detection technologies (Partnership on AI, 2020).

307 https://citizenevidence.amnestyusa.org/
308 https://deeptracelabs.com/
309 https://faculty.ai/
310 https://weverify.eu/
311 https://aiethicsinitiative.org/news/2019/3/12/announcing-the-winners-of-the-ai-and-the-news-open-challenge

New forms of manual and automatic forensics include approaches that build on existing understanding of how to detect image manipulation and copy-paste-splice, as well as evolved approaches customised to deepfakes such as using spectral analysis to spot distinctive characteristics of synthesised speech[312], or the idea of using biological indicators[313] to look for inconsistencies in deepfakes (AlBadawy et al., 2019). A set of approaches has also been proposed to create a so-called 'soft biometric' of key public figures such as 2020 U.S. presidential candidates that will check in a suspected deepfake whether audio and lip movements have been simulated (Agarwal & Farid, 2019; Beavers, 2019). In authentic content there should be a correlation between what the person says and how they say it (a characteristic pattern of head movements related to how that known individuals says particular words).

Other approaches look for physical integrity ('does it break the laws of physics?') issues such as ensuring there is no inconsistency in lighting, reflection and audio, as well reviewing the semantic integrity of scenes ('does it make sense?'), considering audio forensics[314] approaches to identifying forgeries, and identifying image provenance and origins (Moreira, et al., 2018).

Other automated approaches to tracking deepfakes relate to existing automated content detection systems on platforms, including image phylogeny and image provenance based approaches. Image provenance approaches relate most closely to existing image search engines that utilise reverse-image search or other similarity searches to identify previous or similar versions of an image. Image phylogeny approaches draw on similar indexes of existing images to look for the history of image elements and to detect re-use of elements within the frame.

Tools for automated detection of AI-generated text include Grover[315] (Zeller et al., 2019) or the Glitr model[316] (Strobelt & Gehrmann, 2019). Grover is both a generative system as well as a detection system and like other deep learning-based approaches these tools are generally less robust when applied to text generated with different models and datas from those on which they were trained. Early developers of methods and datasets in this area - e.g. Open AI's GPT-2 - (Solaiman et al., 2019) have continued to release information on their code and model weights to facilitate detection of the outputs of GPT-2 derived models. Commercial actors working on anti-disinformation efforts and investigation efforts (as noted in 4.2) are investigating their utility for detecting automatically generated text (Rahman et al., 2019).

6.2.4 Who are the primary actors and who funds these responses?

Existing Internet companies through their commercial models (e.g. targeted advertising) support internal responses as well as some provision of services to third parties. These services include proprietary resources such as automated detection of bots, restricted resources such as information for third-party fact-checkers, and datasets for deepfakes

312 https://www.researchgate.net/publication/333393640_Detecting_AI-Synthesized_Speech_Using_Bispectral_Analysis
313 https://www.researchgate.net/publication/333393640_Detecting_AI-Synthesized_Speech_Using_Bispectral_Analysis
314 https://newsinitiative.withgoogle.com/dnifund/dni-projects/digger-deepfake-detection/
315 https://grover.allenai.org/
316 http://gltr.io/

detection. In some cases, there are public-facing capacities such as similarity search or image-search. In general these are not paid services.

Other approaches, particularly for third-party tools, are a mix of government-challenge grant-funded (e.g. DARPA and EU funds for detection and verification approaches) as well as non-profit initiatives and start-ups.

6.2.5 Response Case Study: COVID-19 Disinformation

One key technical and algorithmic consequence of the COVID-19 pandemic is the move to more automated content moderation and a greater described tolerance for false positives by the major internet companies. Although driven by issues of workplace health and information security as workforces (staff and contracted) move to working remotely, this provides an experiment in a more automated process of content review. Facebook notes that "with a reduced and remote workforce, we will now rely more on our automated systems to detect and remove violating content and disable accounts. As a result, we expect to make more mistakes, and reviews will take longer than normal". They also note that 'normally when we remove content, we offer the person who posted it the option to request that we review the content again if they think we made a mistake. Now, given our reduced workforce, we'll give people the option to tell us that they disagree with our decision and we'll monitor that feedback to improve our accuracy, but we likely won't review content a second time."[317] Other companies are also direct about the consequences of a shift to more automation. Google notes "our automated systems may not always accurately classify content for removal, and human review of these decisions may be slower"[318]. Twitter states that it is: "**Increasing our use of machine learning and automation** to take a wide range of actions on potentially abusive and manipulative content. We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes."[319] YouTube notes that "automated systems will start removing some content without human review, so we can continue to act quickly to remove violative content and protect our ecosystem, while we have workplace protections in place... As we do this, users and creators may see increased video removals, including some videos that may not violate policies."[320]

One study in mid 2020 indicated how difficult it is for Facebook to deal with prolific levels of health disinformation on the site, arguing that the company needs to improve its algorithmic responses (Avaaz 2020). In particular, the study found that only 16% of content identified by researchers as health-related misinfomation carried a warning label. False and misleading health content was viewed 3.8 billion times in the preceding 12 months, peaking during the Covid-19 pandemic, according to the research (Avaaz 2020).

Two risks of automated content moderation are starkly revealed - that in the absence of related human review, it creates ongoing false positives for content policy violations, and that a right to appeal decisions is essential. One observer comments: "With many human content moderators suddenly out of commission, platforms have been forced to acknowledge the very real limits of their technology... Content moderation at scale

317 https://about.fb.com/news/2020/03/coronavirus/
318 https://blog.google/inside-google/company-announcements/update-extended-workforce-covid-19
319 https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html
320 https://youtube-creators.googleblog.com/2020/03/protecting-our-extended-workforce-and.html

is impossible to perform perfectly - platforms have to make millions of decisions a day and cannot get it right in every instance. Because error is inevitable, content moderation system design requires choosing which kinds of errors the system will err on the side of making. In the context of the pandemic, when the WHO has declared an "infodemic" and human content moderators simply *cannot* go to work, platforms have chosen to err on the side of false positives and remove more content." (Douek, 2020).

The companies' statements acknowledge that currently automated systems are not a replacement for human oversight, and this reinforces the need for a robust corrections and appeals systems, as has been highlighted by the UN Special Rapporteur on promotion and protection of the right to freedom of opinion and expression (Kaye, 2018). The same observer cited above further notes: "Content moderation during this pandemic is an exaggerated version of content moderation all the time: Platforms are balancing various interests when they write their rules, and they are making consequential choices about error preference when they enforce them. Platforms' uncharacteristic (if still too limited) transparency around these choices in the context of the pandemic should be welcomed - but needs to be expanded on in the future. These kinds of choices should not be made in the shadows." (Douek, 2020).

6.2.6 How are technical and algorithmic responses evaluated?

The lack of data availability impedes external scrutiny of the inputs, models and outputs of most internal algorithmic processes within platforms. This also has the impact of reducing the public's capacity to evaluate external and third-party algorithms, as outsiders do not have access to either all data within a specific platform, or contextually relevant data around a phenomena to be studied or identified. Nor do members of the public have access to cross-platform data to adequately track disinformation. Both these factors impede effective evaluation.

As noted above, the absence of deeper transparency on usage of algorithmic systems, or on implementation of human rights due diligence prevents effective external evaluation of their effectiveness in countering disinformation or their impact on freedom of expression and other rights (see Llansó et al., 2020; Gorwa et al., 2020). Transparency reports provide aggregate figures on enforcement around for example, false accounts[321], but do not provide detail.

Deepfakes detection models - both forensic and deep learning based - are evaluated against benchmark standards and a test set of similar images that are not part of the training data, but are currently untested in the context of widespread usage 'in the wild' of deepfake or synthetic media imagery created with a wide range of existing and novel approaches.

6.2.7 Challenges and opportunities

For Internet companies, machine-learning enabled approaches to identifying and controlling disinformation benefit from the potential to implement them at scale and at a speed closer to real time than human oversight. They can provide a mechanism for triage of content and for providing insight to humans within the significant teams within companies who hold designated threat responses roles, as well as the large (in-house

321 https://transparency.facebook.com/community-standards-enforcement#fake-accounts

and outsourced) content moderation teams (an estimated 15,000 as of March 2019 at Facebook) (Newton, 2019b). Both these goals may not necessarily align with societally desirable freedom of expression outcomes.

As algorithmic responses, they are subject to both potential implicit and explicit bias in their design and in the training data that is used to develop them (see further discussion below). However, at a specific content item level they are less susceptible to pressure by states and others on individual human operators within a company to take action on a case of claimed disinformation.

For third-parties including fact-checkers, journalists and other investigators, machine-learned enabled tools provide additional mechanisms for understanding content and speeding-up decision-making, however subject to the limitations of not having additional context that is available to the platform companies. These tools may also be used to analyse misapplied or poorly applied platform automated measures and assess impact on freedom of expression.

Current tools, however, are not suitable for identifying disinformation at scale, in real-time, and with very high accuracy. Algorithmic responses within platforms suffer from a range of freedom-of-expression compromising characteristics. Some of these are procedural, some due to the limits of the technical parameters of the AI systems, and others are decisions taken for proprietary reasons or to protect systems from adversarial attack (see Duarte & Llansó, 2017; Llansó et al., 2020). Some are also functions of policies that lack consideration of the international standards for freedom of expression in terms of how they make judgements on potential harm and proportionality. A further complication is when policies are either more vague or more broad than international human rights law in their definition of terrorism, hate speech and incitement to harm (Article 19, 2018a; Article 19, 2018b).

AI-based approaches suffer from the so-called bias problem which occurs at multiple stages of designing, building and implementing an automated system (Hao, 2019). They include problems of how a problem is framed (for example, definitions of what is considered disinformation or inclusion of human rights standards), at the level of data collection when training data may be collected that is unrepresentative, poorly labelled or inadequate (or contain implicit or explicit bias towards a particular group as is the case with AI in other settings), and at the level of preparing the data to ensure the algorithm is focused on the salient characteristics to the objective of the automated system.

Most tools work best when they are trained and applied in specific domains and cannot necessarily be applied with the same reliability across divergent contexts. AI-based systems do not translate well between diverse contexts, particularly when there is inadequate appropriate training data to train the machine learning models. This can result in compromised effectiveness for particular types of content - for example, content from minority populations or languages where there is inadequate or poorly sourced data, as has been the case with assessing the effectiveness of identification of hate speech in Burmese language (Stecklow, 2018) - or over-targeting of particular types of content. Already-marginalised populations face further marginalisation from automated systems. These issues, particularly in terms of understanding less visible communities and less prominent languages have implications for AI systems that analyse discourse in cases where these are applied to disinformation detection.

Additionally all AI and algorithmic systems are designed and implemented with policy objectives in mind that to a greater or lesser extent may align with freedom of expression considerations or may hold implicit or explicit bias at the policy design and framing level.

For example Facebook highlights five values that it uses in its Community Standards (Bickert, 2019). These include Voice, Authenticity, Safety, Privacy and Dignity, and although they do make reference to using 'international human rights standards' to make judgments on cases they do not provide granular detail on how this is done. Rather than implicit bias in the design of an algorithm, internet companies make explicit decisions in their policies around how they understand freedom of expression, with cascading implications into the design and application of algorithms and other automated systems, and in decision-making around what is escalated to human review.

Defining terms is also a challenge with training machine learning systems - given the challenges in defining disinformation (and misinformation) and disagreement between humans on definitions, this lack of precision inhibits building strong data sets. In the cognate field of hate speech, when people are asked to annotate racial slurs, they have been found to agree with each other in only 69% of the cases (Bartlett et al., 2014). The task of distinguishing polite from impolite tweets has been found easier for humans, with agreement ranging from 80% to 95% depending on the language of the tweet (Theocharis et al., 2016). Similarly, the 0-day subsequent performance of deepfake detectors against a novel forgery technique will always be compromised, particularly as long as detection models do not generalise well to new forgery techniques. Deepfake detectors also face significant weaknesses in terms of dealing with the compression and transcoding common to social networks, as well as dealing with adversarial perturbations that disrupt computer vision. There is also significant discussion about how best to present the data derived from arrays of in detectors of forensic manipulation in a human-readable and human-explainable format (Verdoliva, 2020).

In the realm of disinformation, between fact and fabrication, a distinction can be made, but whether the first constitutes truth and the second is always falsehood (as distinct from satire or fiction, or of as yet unknown status) is a lot more complex. This makes automation challenging in regard to this particular area / dimension of content, and likewise with the correlation of content to fake identity and inauthentic behaviour (co-ordinated or not). Audiovisual content also complicates even the first distinction where much content used in disinformation is recycled or mis-contextualised authentic content wherein the underlying content can be factual or truthful but the framing fabricated.

In addition, many machine-learning systems dependent on neural networks - for example, many of the tools for detecting deepfakes and other synthetic media as well as for more effective detection of existing media manipulations - exist in a continuous adversarial dynamic with actors trying to fool them (Verdoliva, 2020).

Although platforms do not provide detailed information on the effectiveness of their automated detection tools, we can learn from state-of-the-art methods about levels of precision in NLP and other areas. As an example, in academic research, state-of-the-art methods for hate speech detection currently have 65-70% precision compared to human detection using the same definition and data set (Wulczyn, Thain, & Dixon, 2017). However, it is hard to give a consistent figure as datasets and tasks vary widely - the highest rates noted in recent studies range up to 92% accuracy (MacAvaney et al., 2019). Even though the Internet companies have access to additional, non-public information about a given post (e.g. its originating IP address), the algorithms are still not sufficiently accurate to be used in a fully automated manner. For instance, recently Facebook's hate speech detection algorithms were triggered by part of the U.S. Declaration of Independence, which resulted in the post concerned being automatically withheld from initial publication (MacGuill, 2018). Even rates of failure of 10% will be magnified rapidly given the scale of content items in any given social network, and also automated systems often combine multiple algorithms with a consequence that mistakes can be magnified

rapidly. If there are serious challenges to identifying hate speech references through machine learning, the fraught issue of automated assessment of disinformation (even on topics like climate change), is even more complicated.

There are implications of these accuracy constraints, and of when (1) false disinformation is wrongly labelled as true or bot accounts are wrongly identified as human; and (2) false positives. Correlatively, there are issues when correct information is wrongly labelled as disinformation or genuine users are wrongly identified as bots. The conclusion is that current automated tools are not suited for independent operation without human oversight or redress possibility.

This is especially true as current automated systems on platforms have procedural weaknesses. These include a lack of oversight and transparency around algorithms, including an inability for independent outsiders to audit where there is bias in design, training data or implementation, or to evaluate the effectiveness of the approach (Ranking Digital Rights 2020). This problem is also noted above in relation to evaluating the approaches for message, actor and behaviour analysis that the companies are implementing.

This lack of transparency also means that erroneous deletion or down-ranking of content or actors combines with a lack of explainability on individual and group decisions to classify content as fitting within a category, such as disinformation. Even attempts to address content moderation with more independent oversight (for example Facebook's Oversight Board, 2019e) do not include the power to change underlying algorithms. Similarly the 'blackbox' absence of algorithmic transparency or explainability impedes usefulness to journalists/fact-checkers when it comes to explaining content decisions.

Bearing this in mind a range of principles for increased transparency exist - including the Santa Clara Principles[322] focused on numbers, notice and appeal. There are also the recommendations of the UN Special Rapporteur David Kaye (Kaye, 2018) on accountability for the Internet companies as well as "transparency initiatives that explain the impact of automation, human moderation and user or trusted flagging on terms of service actions."

One key underlying challenge is that internet platforms use content recommendation algorithms that reinforce related problems of extremism and in-group consolidation of beliefs (Lewis, 2018) and work at cross-purposes or counterproductively to the efforts to challenge disinformation.

For third-party tools, a recent German Marshall Fund report looked at 13 start-ups that aim to use artificial intelligence (and/or machine learning) to fight disinformation. Its top-level findings state that "natural language processing alone can't identify all forms of fakery, and such technology would likely hit several hurdles before ever being implemented." (Schiffrin & Goodman, 2019). Independent tools based on machine learning and seeking to do network analysis face not only the hurdles noted above, but additional ones to platform-based tools, particularly if they must interact with limited data from social media and search sites. An additional hurdle is there is no shared API access or consolidated data between Internet companies. This challenges third-parties as disinformation does not remain on one commercial property but moves between them, as well as across non-commercial platforms, messaging apps, search and video-sharing, and so it is harder to effectively gather cross-platform data on movement and activity around disinformation.

322 https://www.santaclaraprinciples.org/

Other weaknesses specific to third party tools include that, in addition to reliable training data sets from the Internet companies, there are - because of privacy and consent constraints - limited available datasets 'in the wild'. In addition, third-party tools, just like platform-based tools, also exist in an adversarial dynamic with disinformation actors. Algorithmic solutions trained on previous data that have not been re-trained or updated will likely miss new forms of misinformation and disinformation, with significantly worse performance.

Limitations of image search, similarity search, media forensics and image phylogeny tools

Current reverse image and related image similarity using search engines offer generally good accuracy. However, they do depend on the exhaustiveness of the indexing done by the search engines in order to identify prior images, and there is an absence of robust reverse video search that is effective for video modification and edits. If more fake images are indexed than the original, it may become difficult to retrieve the original image or video, especially over time. Reverse video search is computationally complex and currently not publicly available on platforms.

In addition, there are technical gaps in terms of media forensics tools. Most do not function well with compressed media, with low-resolution media, or provide easily human-readable information. Combined with a significant deficiency in media forensics understanding among journalists, media and fact-checkers, advances in media forensics tools are not always well-aligned with the needs of civil society and media needs (see Gregory & French, 2019). These deficiencies include addressing the issues of media quality and compression, and the need to make decisions rapidly and to explain them to the sceptical public.

In conclusion, there are still significant policy, practical and ethical barriers to more widespread usage of AI and machine learning systems for message, actor and activity detection at scale, in terms of their relation to freedom of expression, accuracy, impact on vulnerable populations and transparency/capacity for appeal. They are not suitable for usage beyond in a semi-automated and assistive capacity. Tools for single content item evaluation - for example to confirm conventional non-AI forensic manipulation in a photo - are more robust, yet they also face data gaps and gaps in the capacity of journalists and others to utilise them.

Despite improvements in overall human rights due diligence within within policies by internet, search and messaging companies[323] (see Ranking Digital Rights 2019), important gaps still remain (Hogan, 2018). These issues have elicited criticism for failure to systematically invest in impact assessments that thoroughly engage with civil society and other stakeholders as the companies enter new markets/societies with existing products. Similarly, the companies are criticised for not evaluating emerging risks in existing markets (the Facebook post-hoc assessment of its impact in Myanmar is a publicised exception in response to civil society critiques, Facebook 2018b). There is a lack of transparency which complicates external oversight on platforms and their algorithms, including access to better evaluation data on successful identification as well as identified false positives and false negatives. Additionally, the companies are criticised for not engaging in "abusability testing", where "platforms invest resources into seeing how their platforms can be abused to harm consumers. I think that smart policy would incentivise that kind of investment, as we have seen that kind of incentivising around cyber security in the last 10 years" (Soltani,

[323] https://rankingdigitalrights.org/index2019/indicators/g4/

2018). Similarly, there is discontent about an apparent absence of 'freedom of expression by design approaches' (Llansó et al., 2020).

Gaps in data include - as noted above - absence of real-time, cross-platform and cross-company data for researchers and journalists to enable better detection.

There are specific gaps in relation to tools for authentication of audiovisual media content including reverse video search and robust similarity search in platforms and messaging tools, as well as improved provenance tools that provide opt-in machine-readable and human-readable signals. In addition, better tools are needed for analysing memes as disinformation (see Theisen et al., 2020), and for distinguishing across multiple elements of a media item between satire and disinformation.

As deepfakes and synthetic media become more widely available, there is a need to built on shared training datasets (generated and new forgery approaches identified 'in the wild'), generalisable to new forms of falsification and to the extent possible, given adversarial dynamics, accessible to a range of users with explainable results (Leibowicz, 2019). As multiple indicators will be needed across a range of manipulations, so dashboards and detector tools will need to combine multiple forensic and content authentication tests into human-readable formats, useful to journalists and investigators (Verdoliva, 2020). This will need to be complemented by investments in forensics capacity within the journalistic and investigatory worlds to interpret new forms of machine-learning based image manipulation.

6.2.8 Recommendations for technical and algorithmic responses

Given the challenges and opportunities identified above, and the considerable freedom of expression implications of algorithmic responses, the following policy recommendations can be made.

International organisations and States could:

- Invest in monitoring, measuring and assessing the impacts of technical responses to disinformation against human rights frameworks.
- Support the development of independent initiatives that embed impact measurement and evaluation to increase knowledge about the efficacy of technical responses, ensuring that transparency and verifiable criteria are involved.
- Work with internet communications companies to ensure the responses that they initiate are appropriately transparent and measurable, as well as implemented on a truly global scale.
- Encourage the companies to co-operate transparently across basic norms, and produce comparable data that can be used to develop an overview of the problem across different services and related policy frameworks.
- Support initiatives towards ensuring privacy-preserving, and equitable access to key data from internet communications companies, to enable independent research and evaluation on a truly global scale into the way algorithmic responses impact on the incidence, spread and impact of online disinformation.

- Consider implementation of independent national ombuds facilities to help give users recourse to independent arbitration with respect to appeals for unfair automatic content removals and account suspensions.

Internet communications companies could:

- Support independently managed funds for independent research and evaluation of the effectiveness of companies' algorithmic responses to disinformation.
- Work together to improve their technological abilities to detect and curtail disinformation more effectively, and share data about this, as disinformation often exploits cross-platform methods.
- Recognise the limits of automation in content moderation and curation, and expand the human review as well as appeals process.
- Produce detailed public transparency reports, including details on automated removals of disinformation and suspension of accounts spreading disinformation, as these responses can have significant human rights and freedom of expression impacts.
- Reassess how the technology of current business models facilitates the efforts of those producing and distributing disinformation (such as in ranking and recommendations), and how this may undercut other technical efforts to identify and act against disinformation.

Civil society organisations and researchers could:

- Continue independent monitoring and evaluating the successes and dangers of technical and algorithmic responses developed by internet communications companies.
- Study the technological dimensions of cross-platform disinformation campaigns to get a more rounded, holistic perspective on the problem and responses to it.
- Work towards developing new tools to assist journalists, news organisations and other verification professionals with efficient detection and analysis of disinformation, as well as with the crafting and effective promotion of debunks and authoritative information.
- Reinforce trustworthiness and transparency in regard to their roles in technological responses to tackling disinformation.

6.3 Demonetisation and advertising-linked responses

Author: Kalina Bontcheva

Economic responses to disinformation include steps designed to stop monetisation and profit from disinformation and thus disincentivise the creation of clickbait, counterfeit news sites, and other kinds of for-profit disinformation. Demonetisation responses can also target misleading or false content that is created for purposes other than profiteering alone, including when this is fused with hate speech (while demonetisation can be applied to stand-alone hate). The StopHateForProfit campaign of 2020 seeks to apply demonetisation to the package of "hate, bigotry, racism, antisemitism, and disinformation".[324] However,his section will survey this kind of economic responses which are aimed specifically at disrupting the advertising-based monetisation of online disinformation (e.g. making false news sites non-viable).

It must be noted that this section will cover only the economic aspects of online advertising (based on making money off disinformation by attracting advertising through automated systems) and how internet companies try to disrupt these through current measures. This should be distinguished from the primarily political motives for disinformation spread through voter-targeted advertising during elections, which will be addressed in Section 5.3. At the same time, this chapter includes consideration of responses to those actors who directly seek returns from placing advertisements which themselves include disinformation. By acting against such adverts, the Internet communications companies disincentivise such activity. In this sense, demonetisation in this chapter refers to (i) preventing the placement of adverts next to disinformational content, and (ii) prevention of adverts that contain disinformation from appearing/ remaining on the company's service.

6.3.1 What and who do demonetisation and advertising-linked responses target?

Through disinformation, traffic is driven to websites where online advertising can be used for monetisation. This traffic is stimulated through a combination of clickbait posts and promoted posts, i.e. adverts (which themselves could be clickbait in nature). There are numerous false news sites and fabricated online profiles (e.g. on Twitter, Facebook) and groups, which are created as part of this process. To give just one example, a man created and ran, in a coordinated fashion, over 700 Facebook profiles (Silverman, 2017a), promoting links and attracting clicks to false content on websites, which in turn generated revenues from the advertising displayed alongside (Silverman, 2016). Other examples include Google Adsense and doubleclick being used to fund the Suavelos network of deceptive white supremacist websites in France (EUDL, 2019c) and an Africa-based network of for-profit junk media outlets and clickbait websites, which was publishing health disinformation and which also directly copied articles from particular media outlets to make it seem legitimate (EUDL, 2020).

[324] https://www.stophateforprofit.org/

A clickbait post is designed to provoke an emotional response in its readers, e.g. surprise, intrigue, thrill, humour, anger, compassion, sadness, and thus stimulate further engagement by nudging readers to follow the link to the webpage, which in turn generates ad views and revenues for the website owner. Clickbait typically omits key information about the linked content (Chakraborty et al., 2017), in order to create a curiosity gap (Loewenstein, 1994) and thus entice users to click. This by definition often implies that clickbait is not an accurate representation of the content it promises, and can contain disinformation as false or misleading content. The sensationalist and emotive nature of social media clickbait has been likened to tabloid journalism and found to provide an "alternative public sphere for users drifting away from traditional news" (Chakraborty et al., 2017). Clickbait tweets, for example, have been found to retain their popularity for longer, and attract more engagement, as compared to non-clickbait tweets (Chakraborty et al., 2017). These characteristics make them highly successful in propagating organically online mis- and disinformation through networks of genuine users, as well as being used in many highly-viewed adverts. Clickbait may be within direct content or as an ingredient in advertising.

Online advertising is a common means towards monetising deceptive and false content on junk news sites, as the creators receive payments when adverts are shown alongside the junk content. For instance, when adverts (often from major brands) were shown on YouTube at the start of videos containing health misinformation, this generated revenue both for the platform's owner (Google) and the publisher of the videos on fake cancer cures (Carmichael & Gragnani, 2019). Creators of fake sites and videos have claimed to earn between $10,000 and $30,000 per month from online advertising, e.g. the CEO of Disinfomedia (Sydell, 2016).

A particularly effective type of online adverts are the so called 'dark ads', which are only visible to the users that are being targeted (e.g. voters in a marginal UK constituency (Cadwalladr, 2017)) and do not appear on the advertiser's timeline. They have been used during political campaigns to spread disinformation, with the intent of influencing voter outcomes (Cadwalladr, 2018). Moreover, due to their highly personalised nature, dark ads can be used to target susceptible users with disinformation which they are likely to believe is correct. As dark ads are hidden from view of other users, disinformation within cannot be discussed or counter-evidence posted by the user's friends.

Facebook adverts, including highly targeted 'dark ads', have also been used recently to carry falsehoods and sell fake products, using inter alia videos and materials stolen from the popular Kickstarter crowdfunding platform (Bitten, 2019). Another multi-million dollar scam on Facebook used a combination of rented Facebook accounts, deceptive adverts, and subscriptions to defraud less savvy users (typically from the baby boomer generation) (Silverman, 2019).

Other internet communications companies are not immune. For instance, in late 2019 the white supremacist Suavelos network published a false anti-immigrant story on suavelos.eu, which was debunked by fact-checkers AFP[325]. This prompted an in-depth investigation by the EU DisInfo Lab (EUDL, 2019c) which uncovered that the Suavelos network (consisting of several websites, Facebook pages, a YouTube channel, and Twitter and VKontakte accounts) was making money from advertising via Google Adsense or Doubleclick and through related and similar sponsored content using Taboola.

[325] https://twitter.com/AfpFactuel/status/1155125308535840768?s=20

Promoted posts on Facebook and Twitter are marked as advertisements and can be reposted, liked, replied to, etc. as any normal post can. Advertisers are billed by the social platform based on the amount of engagement generated, i.e. likes, shares, clicks and views.

In many cases advertisers can choose which users will see the promoted post, based on information such as geographic location, gender, interests, device type, or other specific characteristics. When adverts are targeted at a very narrow set of users (the so called "dark ads"), with very specific profiles, the practice is called micro-targeting.

As users visit websites and social media platforms, they are willingly or unwittingly giving away invaluable personal information, e.g. their location, mobile device used, IP address, browsing history, time spent on particular content while scrolling, social media engagements (e.g. likes and shares), and mood (emoticons, gifs). Social profiles are typically data rich and include further personal data, including birthday, relationship status, family members, workplace, education history, etc. Moreover, users' online behaviour is continuously tracked through technology such as cookies, tracking scripts and images, display adverts, and CSS/HTML code. All this data is what enables the automated profiling of users and the resulting micro-targeted delivery of personalised advertising and/or content.

Because of inter alia the instrumentalisation of these targeting powers for spreading falsehoods, many policy makers have called for transparency and regulation of online advertising as important steps towards disrupting monetisation of online disinformation:

> *"Platforms should adapt their advertising policies, including adhering to "follow-the-money" principle, whilst preventing incentives that lead to disinformation, such as to discourage the dissemination and amplification of disinformation for profit. These policies must be based on clear, transparent, and non-discriminatory criteria (Buning et al., 2018)."*

6.3.2 Who do demonetisation and advertising-linked responses try to help?

Demonetisation responses try firstly and foremostly to limit the circulation of for-profit online disinformation and thus protect citizens from fraudulent products, harmful "miracle cures", and political disinformation during elections and referenda. It is unclear to what extent other or particularly "white-listed" content could be promoted for the purposes of attracting advertising, and there are issues around the practice of allowing advertisers to blacklist (and therefore avoid) placement next to certain content - such as blacklisting association with any COVID-19 content (whether true or false).[326]

Secondly, ad screening and ad transparency measures are being implemented in part by the internet companies, in order to protect their multi-billion ad revenues, as advertising increasingly moves online and becomes automated (WARC, 2019). Complaints by users and campaign advocacy have led to major advertisers withdrawing patronage because of juxtaposition next to hate-speech.[327]

[326] See https://gfmd.info/press-release-emergency-appeal-for-journalism-and-media-support/
[327] https://www.businessinsider.fr/us/facebook-fbrape-ad-boycott-2013-5

A key assumption behind economic responses is that internet and media companies have significant power to control and prevent the monetisation of disinformation through their services. Secondly, it is assumed that the companies' business models and associated "attention economics" are not intrinsically favourable to disinformation, and that the captains of these enterprises are willing to invest time and effort to implement and enforce such responses.

The successful implementation of these responses relies on the companies' social and ethical responsibility and their ability to detect and demonetise effectively for-profit disinformation. Due to the sheer volume of promoted posts and adverts on these companies' services, economic responses are resorting primarily to algorithmic automation[328] with the assumption that this is sufficiently sophisticated to detect and determine the course of action for disinformation as regards monetisation dimensions. Only in some cases are reported adverts/promoted posts subject to manual screening. However, this is not always effective.[329] This can be further problematic for two reasons. Firstly, there needs to be adequate provision for redress for content wrongly removed under these means. Secondly, in order for the adverts review process to be triggered, users need to report the adverts first. It is currently unclear, however, whether the majority of users (especially children and adults over 50) are aware that they can do so. On some platforms users can also find out why they are being shown a given ad and indicate if they wish to stop seeing adverts from a particular advertiser. However, more evidence is needed that users are aware of this potential, where it is offered, and are therefore making active use of it.

6.3.3 What output do demonetisation and advertising-linked responses publish?

The report of the EU High Level Expert Group on disinformation (Buning et al., 2018), government reports (e.g. (DCMS report, 2018c)) and independent fact-checking organisations (e.g. (FullFact, 2018)) have strongly advocated that all paid-for political and issue-based advertising data must be made publicly accessible for research by the internet communications companies hosting the adverts. This includes detailed information about the advertising organisation, country of origin, and at whom the adverts are targeted. Details on the current implementation of ad transparency by internet communications companies was discussed already in Chapter 6.1 in the context of curatorial responses.

Overall, ad transparency libraries are a key element of enabling independent scrutiny not only of political advertising, but also of the economic responses implemented by internet communications companies with the aim of limiting the promotion and monetisation of disinformation through online advertising.

At present, however, their reach and utility are insufficient, not only in terms of geographical coverage, but also in terms of ad topics. For instance, except for Facebook, all other ad libraries currently do not provide transparency information on COVID-19 and related adverts, since this is not one of the issues included in their issue ad scope. This significantly impedes independent scrutiny of the extent of removed COVID-19 adverts.

328 https://en-gb.facebook.com/business/help/162606073801742
329 https://www.consumerreports.org/social-media/facebook-approved-ads-with-coronavirus-misinformation/

As discussed in Chapter 6.1, there is also limited information in the transparency reports published by the internet communications companies with respect to demonetisation of websites and accounts spreading disinformation.

6.3.4 Who are the primary actors and who funds these responses?

Demonetisation efforts are self-regulatory measures being implemented by the internet communications companies in response to pressure from national and international governing bodies and policy makers. Examples of regulatory and co-regulatory measures towards ensuring transparency of demonetisation and online advertising include the U.S. Honest Ads Act (Warner, 2017) and the European Commission's Code of Practice (European Commission, 2018c). The latter seeks to involve internet communications companies (Google, Twitter, Facebook, Microsoft, and Mozilla), advertisers, and the advertising industry. Further details on legislative, pre-legislative, and regulatory responses are provided in Section 5.1.

As a result, many internet communications companies (using their own resources) have been taking steps towards disincentivising the production of disinformation for financial gain (including control over online adverts). Similar to the situation with curatorial responses (see Chapter 6.1), reliable figures on platform expenditure on demonetisation efforts are hard to come by. A high level, comparative overview of demonetisation and ad screening measures across 9 internet communication companies were discussed in the previous Chapter 6.1. Here we will analyse further ad-oriented measures in particular:

- **Google**: In April 2019 alone[330], Google reported that a total of 35,428 EU-based advertisers violated their misrepresentation policy, with offending adverts across Google Search, YouTube, and third-party websites who display Google adverts for monetisation purposes. However, as Google's policies are wider than demonetisation of disinformation on its own, the impact of these policies specifically on disinformation spread is currently not quantified by the company itself. During the same time period, Google identified, labelled, and made publicly available 56,968 EU-based political adverts from verified advertisers, but at the time of writing does not provide transparency reporting on issue-based adverts. Specifically, there is a need for a quantified report of measures aimed at demonetising disinformation websites, since a recent independent study (Global Disinformation Index, 2019) revealed Google as the ad platform providing 70% of adverts to known disinformation websites, leading to over $86 million in ad revenue for these sites.

- **Facebook**: In the same time period, Facebook[331] took action against 600,000 EU-based adverts containing low quality, false, or misleading content, which violated its policies. Similar to Google, it is unclear how many of these were specifically disinformation demonetisation efforts. Facebook is currently unique in providing transparency information not only on political, but also issue-based adverts under the following categories: Immigration, Political Values, Civil & Social Rights, Security & Foreign Policy, Economy, and Environment. This has enabled some level of independent scrutiny of such adverts, including pro- and anti-vaccine adverts (Jamison et al., 2019). In January 2020, it was announced

330 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59226
331 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59225

that users will be able to reduce the number of political and social issue adverts they see on Facebook/Instagram (Leathern, 2020; Nuñez, 2020). However key questions[332] are being raised over Facebook's policy (Leathern, 2020) of not following Google's policy of restricting political advert targeting and not screening political adverts for disinformation, such as allowing fabricated climate change-related political advertising to run on the platform (Kahn, 2019); and promoting adverts containing statements rated false by fact-checkers during the election in Sri Lanka (Wong, 2019b). This is an area where Facebook could decide to refer to their new oversight board[333], in addition to this body's stated remit of reviewing the company's decisions on issues of removal of content. Two other areas in need of further attention are for-profit Facebook account rental (Silverman, 2019) and small-group and individual Facebook-based fundraising campaigns, which are successfully promoting anti-vaccination messages (and other contentious social issues) in violation of the platform's policies (Zadrozny, 2019).

- **Twitter**: Between January and March 2019[334], Twitter rejected EU-based 4,590 adverts for violating its Unacceptable Businesses Practice policy and another 7,533 EU-based adverts for non-compliance with its Quality Ads policy. It is unclear again how many of these were specifically disinformation. As of November 2019, Twitter banned political adverts globally.[335]

- **YouTube** has received $15 billion in advertising revenue in 2019 alone (Statt, 2020). In general, YouTube video creators receive 55% of the revenue when an ad is shown before or during their video, with the remaining 45% being retained by YouTube as advertising revenue (Tameez, 2020). The Google-owned service has policies[336] on how YouTube channels can monetise content by earning revenues from ad placement. When videos and channels are found to violate these policies, they can be demonetised or removed. In some cases this has led to self-censorship by content creators for fear of being demonetised (Alexander, 2020), as well as accusations of disparities in the way videos from premium-tier content creators are treated as compared to those from regular content creators (Alexander, 2020). Concerns have been raised by users who were mistakenly demonetised by YouTube about the lack of transparency of YouTube's decision, lack of provision of an effective appeals mechanism, and no options being provided for recovery of lost ad income (Goggin & Tenbarge, 2019). In addition, in January 2020 an independent study showed that despite YouTube's stated policies and efforts, adverts paid for by the top 100 brands were funding climate misinformation (Hern, 2020). The affected brands were not aware that their adverts were shown before and during videos containing misinformation.

- **Reddit**[337]: As of 15 May 2020, the company only accepts U.S.-based advertisers and adverts and all of these undergo manual review. In a novel approach, political adverts will have their user comments enabled for at least 24 hrs, and advertisers are strongly encouraged to engage with the users and their comments. There is also a new political adverts transparency subreddit.[338]

332 https://www.forbes.com/sites/mnunez/2020/01/09/facebook-will-let-you-reduce-the-number-of-political-ads-you-see---but-it-still-wont-stop-politicians-from-lying
333 https://www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html
334 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59227
335 https://twitter.com/jack/status/1189634360472829952
336 https://support.google.com/youtube/answer/1311392?hl=en-GB
337 https://www.reddit.com/r/announcements/comments/g0s6tn/changes_to_reddits_political_ads_policy/
338 https://www.reddit.com/r/RedditPoliticalAds/

- **TikTok**[339] has banned election-related, advocacy, and issue-based adverts from the platform. Concerns have been raised (Kozlowska, 2019), however, that TikTok's entertainment-oriented format, its serendipitous discovery recommender algorithms, and its relative lack of preparedness to detect and contain disinformation are being exploited to spread and promote political campaign messages, conspiracy theories, and pseudoscience.

Automated ad brokerage and exchange networks[340] buy and sell web advertising automatically, which in 2019 was estimated as being worth U.S.$84bn or 65% of digital media adverts (WARC, 2018). The main target markets are the United States, Canada, the United Kingdom, China, and Denmark (WARC, 2018). Major operators include Google[341], The Rubicon Project[342], OpenX[343], AppNexus[344], Criteo[345]. Among them, available sources suggest that only Google has so far committed to providing some degree of ad transparency and only in relation to political adverts. This however is still susceptible to being seen as insufficient, since disinformation websites and deceptive adverts often monetise through purely economic scams, e.g, `free' product trials (Silverman, 2019). At the same time, a September 2019 independent analysis (Global Disinformation Index, 2019) of programmatic advertising on 20,000 disinformation domains concluded that they monetise unhindered over U.S.$ 235 million through ad exchanges. The highest market share was found to belong to Google, which accounted also for the highest estimated amount of revenues for these disinformation sites (over U.S.$86 million), followed by AppNexus (over U.S.$59 million), Criteo (over U.S.$53 million), and Amazon (just under U.S.$9 million). Automatic placement of advertising, and matching to certain content, is a feature that can be easily exploited by disinformation producers.

Some advertisers have started recently to withhold adverts from Facebook, Google, Twitter, and other services offered by the internet communications companies, as a way of demonetising these companies and incentivising them to address more thoroughly and reliably disinformation, especially cases when it can incite violence or suppress voting[346]. These boycott measures have already resulted in significant losses.[347] This momentum gained ground during 2020 with the Stop Hate for Profit movement which listed almost 600 participating businesses by mid-year.[348]

Journalists, **civil society and media organisations**, **fact-checkers and scientists** are also key actors who uncover online scams that harness or profit from disinformation; and also monitor, evaluate, and advise on the implementation of economic responses aimed at demonetising disinformation. Given the largely voluntary, self-regulatory nature of company-implemented economic responses, the role of these independent actors has been both essential and also significant.

339 https://newsroom.tiktok.com/en-us/understanding-our-policies-around-paid-ads
340 https://digiday.com/media/what-is-an-ad-exchange/
341 https://marketingplatform.google.com
342 https://rubiconproject.com/
343 https://www.openx.com/
344 https://www.appnexus.com/fr
345 https://www.criteo.com/
346 https://www.bbc.co.uk/news/business-53204072; https://www.bbc.co.uk/news/business-53174260
347 https://www.bloomberg.com/news/articles/2020-06-27/mark-zuckerberg-loses-7-billion-as-companies-drop-facebook-ads
348 https://www.stophateforprofit.org/participating-businesses

6.3.5 Response Case Study: COVID-19 Disinformation

In the context of COVID-19, steps were taken by the internet companies to stop people making money from coronavirus disinformation and thus to try and remove incentives for creating clickbait, counterfeit news sites, and other kinds of for-profit disinformation on this topic.

There have been two main kinds of economic responses so far: advertising bans and demonetisation of false or misleading COVID-19 content.

- While Facebook does not ban disinformation in political adverts, in this case (alongside Google[349]) the company has taken proactive steps to limit COVID-19 disinformation in Facebook and Instagram adverts, as well as reduce economic profiteering from the pandemic.[350] This is through excluding adverts for testing kits, sanitiser, masks and "cures" at inflated prices, often promoted through click-bait disinformation claims. However, due to the automation-based method used for advert screening, rogue advertisers have found ways to get around the ban[351] through synonymous words and hijacking of user accounts. Google and Bing's demonetisation efforts have also been subverted and their search technology still sometimes displays pages that sell dubious COVID-19 related products[352].

- Early on in the pandemic, Google and Twitter also instituted a blanket ban of all adverts that mention coronavirus and COVID-19 except those placed by government entities or other authorised official sources. This led to the unwanted effect of preventing other legitimate entities from launching helpful information campaigns through adverts. As a result, Google lifted the ban in early April 2020.[353] Twitter's position remained unchanged as of early April 2020: "Twitter prohibits all promoted content that refers to COVID-19. The only exceptions to this prohibition are approved Public Service Announcements (PSA's) from government and supranational entities, news outlets that currently hold a political content exemption certification, and some organizations who have a current partnership with the Twitter Policy team."[354]

- Beyond advertising, YouTube[355] has taken measures to ensure ethical monetisation of content mentioning or featuring COVID-19 by requesting all content is fact-checked by its authors and that its guidelines are followed. When violations are detected, the company says it aims to either remove the offending COVID-19-related content, limit its monetisation, or temporarily disable monetisation on the channel, although it does not provide statistics on this issue.

349 https://blog.google/inside-google/company-announcements/covid-19-how-were-continuing-to-help/
350 https://about.fb.com/news/2020/03/coronavirus/#exploitative-tactics
351 https://www.infosecurity-magazine.com/news/ban-hasnt-stopped-covid19/
352 https://searchengineland.com/a-look-at-googles-recent-covid-19-related-policies-in-search-330992
353 https://www.axios.com/google-coronavirus-advertising-6ff1f504-201c-435a-afe5-d89d741713ac.html
354 https://business.twitter.com/en/resources/crisis-communication-for-brands.html
355 https://support.google.com/youtube/answer/9777243?p=covid19_updates

6.3.6 How are demonetisation and advertising-linked responses evaluated?

The EU Commission released an independent assessment of the effectiveness of the Code of Practice on Disinformation (Plasilova et al., 2020), which concluded specifically on demonetisation efforts that:

- **The effectiveness of ad placement measures**: due to lack of sufficiently detailed data, it was not possible to establish the effectiveness of the measures implemented so far by the internet communications companies. The conclusion here was that: "Currently, the Code does not have a high enough public profile to put sufficient pressure for change on platforms. Future iterations of the Code should refer to click-baiting as a tool used in disinformation and specifically ad placements."

- **Transparency of political and issue-based advertising**: the evaluation acknowledged the positive results achieved so far in this area, however adding that there is still significant room for improvement, especially with respect to issue-based advertising.

- **Empowering the research community**: lack of data is still a very significant problem hindering independent research into disinformation and the accurate assessment of the effectiveness of measures implemented by the internet communications companies in reducing the spread of disinformation (and in that context, also specifically the success or otherwise of demonetisation efforts).

The independent evaluation of platform measures (Plasilova et al., 2020) also concluded that: "A mechanism for action in case of non-compliance of the Code's Pillars could be considered. To that effect, the European Commission should consider proposals for co-regulation within which appropriate enforcement mechanisms, sanctions and redress mechanisms should be established." In particular, the need to ensure that economic responses to disinformation were implemented uniformly across all EU Member States was highlighted.

The evaluation (Plasilova et al., 2020) also proposed a number of Key Performance Indicators (KPIs) that need to be implemented. Here we include a selection of those directly relevant to evaluating the success of economic responses to disinformation:

- **Scrutiny of adverts and limiting disinformation within them**: total turnover received by the advertising operators from advertisements placed; total of foregone (lost) revenue due to certain accounts being closed; total advertising revenue from the top 100 websites identified as prominent purveyors of disinformation. Regular monitoring and reporting these KPIs would show over time whether these measures are improving in effectiveness.

- **Transparency of political and issue-based adverts**: proposed KPIs include number of mislabelled political and issue-based adverts; and ratio of total turnover of issue-based advertising with revenue lost due to accounts closed down due to breach of issue-based advertising policies.

A prerequisite for measuring these KPIs is that the companies provide much more granular and thorough information in their ad libraries than is currently the case (Leerssen et al., 2019), including the need to widen very significantly their extremely limited

present geographic reach; go beyond political and include all adverts; improve targeting information provision and advertiser transparency.

Despite the lack of such all encompassing information, media organisations, civil society and independent researchers are nevertheless able to carry out small-scale evaluations and investigations around specific case studies which provide important insights into the present limitations of economic responses to disinformation. Examples include:

- Facebook/Instagram allowing advertisers to micro-target the 78 million users which the platform has classified as interested in "pseudoscience" (Sankin, 2020);
- Cases of forcing authorities to resort to lawsuits due to the platforms' non-adherence to campaign finance laws for political adverts (Sanders, 2020);
- Continued failures to stop the amplification and enforce demonetisation of thriving networks of junk news sites (EU Disinfo Lab, 2020) or accounts violating the site's terms of service (Ingram, 2019: Webwire, 2020:, EU Disinfo Lab, 2019c), despite widely publicised efforts to the contrary;
- Inability to distinguish between legitimate, quality journalism from other content leading to demonetisation and content removal actions that infringe on freedom of expression and the right to information (Taibbi, 2019);
- Inaction towards limiting disinformation and misleading political advertising and its negative impact during elections (Reid & Dotto, 2019; Tidy & Schraer, 2019; Who Targets Me, 2019).

6.3.7 Challenges and opportunities

These economic responses to disinformation, if implemented properly, offer the promise and the opportunity to reduce the creation and propagation of for-profit disinformation.

However, the majority of economic responses are currently largely in the hands of private actors, where inconsistent and opaque decisions are being made. There is insufficient advertising transparency in the information provided by internet communications companies, thereby preventing independent scrutiny by journalists and researchers. The problem is acutely present across many platforms and countries not only for healthcare (e.g. COVID-19) or issue adverts, but also for political adverts.

The patchwork of policies and approaches between different companies reflects pluralism and diversity, but it can hinder an overall effective industry-wide response to demonetising disinformation. It can also conceal both immediate and enduring risks to the rights to freedom of expression and privacy by corporate actors.

These challenges have been brought into sharp focus by the COVID-19 pandemic, which also represents a very significant opportunity for urgent action by the internet communications companies towards ensuring full transparency, accountability, and multi-stakeholder engagement. In this way, these corporations can demonstrate their goodwill beyond the bottom line and their sincere interest in improving policy and practices to support quality information. This could involve a mix of curational policies to ensure upgrading credible news outlets and other recognised authoritative content providers, and downgrading or removing false content on one hand, and demonetisation efforts linked to this.

6.3.8 Recommendations for demonetisation and advertising-linked responses

The challenges and opportunities identified above and their significant implications for freedom of expression give rise to possible recommendations for action in this category of responses.

Internet communications companies could:

- Improve the reach and utility of their advertising transparency databases towards global geographical coverage; inclusion of all advertising topics (not only political ones); and provision of comprehensive machine-readable access, which is needed to support large-scale quantitative analyses and advertising policy evaluations.
- Produce detailed public transparency reports, including specific information on demonetisation of websites and accounts spreading disinformation.
- Implement screening of political adverts for disinformation through harnessing the already established independent fact-checking efforts.
- Enable user comments on adverts, ideally from the moment they are published and for at least 24 hours. This will enable flags to be raised on potentially-harmful content as a precursor to possible further steps.
- Effectively address the problem of 'account rentals' (i.e. paid use of authentic user accounts by disinformation agents) to curtail the practice of individuals' accounts being exploited for money-making through disinformation and related-advertising.
- Work together to improve their ability to detect and curtail monetisation of disinformation, as monetisation often exploits cross-platform methods.

Advertising brokerage and exchange networks could:

- Step up their monitoring of disinformation domains and work in close collaboration with fact-checkers and other independent organisations in implementing efficient, effective, and scalable methods for demonetisation of disinformation websites and content.
- Implement full advertising transparency measures, as per those recommended for internet communications companies.
- Work together to implement a consistent approach to advertising screening and transparency across networks, which could also be used as a way of spreading the cost of advertising quality screening and transparency measures.

Governments and international organisations could:

- Provide ongoing funding for independent monitoring and compliance evaluation of demonetisation efforts implemented by companies and advertising brokerage and exchange networks.
- Negotiate with these commercial actors about ensuring full transparency and access to data as prerequisites of independent oversight of economic self-regulatory responses to disinformation.

- Encourage internet communications companies and advertising exchange networks to implement appropriate responses to disinformation on the basis of electoral laws and freedom of expression norms, and do so in all countries where their services are accessible .

- Strongly encourage and, if required, demand the adoption of quantifiable Key Performance Indicators (KPIs) for independent measurement and assessment of the effectiveness of demonetisation responses to disinformation.