Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression

A Report by Students at Leiden Law School

Written by *Niamh Coghlan, Amalia Flenga, Carola Gioco, Lindsay McCosh, Flavia Mencarini, Guilherme Migliora, Melanie Moulder, Ona Lorda Roure, Uditi Saha, Agnese Soverini, and Ruoxin Su*

with Dr Mark Leiser, Leiden Law School

Purpose: To inform the Special Rapporteur's annual thematic report presented to the Human Rights Council at its 47th session in June 2021

Executive summary:

1. This report identifies the main challenges posed by disinformation. It analyses the impact of measures adopted by States as well as platforms to combat this phenomenon. Based on the findings of our research, the report provides a series of recommendations to the Special Rapporteur.
2. Disinformation has consequences for both individuals and society. Identified issues include population control, loss of consumer protection, as well as increasing democratic deficits and discriminatory practices.
3. States have deployed various strategies to counterbalance the negative effects of disinformation. Some have used 'command-and-criminality' methods to punish those spreading disinformation. Often these come with disproportionate penalties. Others have launched official programmes that include platforms monitoring or using independent organizations to debunk misleading online news and information. If abused, however, these measures can encroach basic human rights like freedom of expression, the right to liberty and security, etc.. These measures can also hamper the work of journalists and the public's right to access information.
4. The report highlights both legal provisions and guidance at the international and regional levels addressing human rights abuses arising from far-reaching and over broad measures countering disinformation.
5. Platforms have developed policies around electoral/civic integrity, COVID-19, manipulated media, impersonation and identity fraud, and fake engagement to limit the spread of disinformation.
6. Many platforms have developed methods to improve the transparency of their measures, while others continue to shield how they regulate disinformation from public view. The most effective measures consider the harm caused by disinformation. Fair measures make room for expression that includes false or misleading information but does not cause harm. Fairness also requires that policies regulating the dissemination of disinformation affect all users equally and that platforms address any discriminatory impact of their policies.
7. The appeals processes offered by platforms range in transparency and effectiveness. Facebook's new Oversight Board is an example of a transparent and external appeals process. Yet its effectiveness is uncertain.
8. Other measures which have been effective in protecting freedom of expression in the fight against disinformation rely on independent and external fact-checking organizations or collaboration with external experts for disinformation regulation.
9. Unfortunately, some measures to address disinformation have aggravated human rights violations.
10. Censorship is not an appropriate method for combatting disinformation. De-platforming users, internet shutdowns, and the blocking of websites are tools for controlling information.
11. To better combat the spread of disinformation and any threats to human rights, the Special Rapporteur should encourage both States and platforms to create better policies surrounding the use of political advertisements, address the connection between hate speech and disinformation, encourage access to effective remedies for individuals harmed by disinformation. Furthermore, platforms that are instrumental in shaping public discourse should submit to independent oversight and independent audit of content removal decisions.

Introduction:

The regulation of disinformation delivered via digitally mediated platforms is at a crossroads. Some have referred to the problem of fake news, disinformation, and online manipulation as a 'crisis-infodemic'.[1] The world over, governments have implemented strict command-and-criminalise measures to combat the problem of fake news and disinformation, while others have proposed new models of regulation that would impose a duty of care on platforms to prevent harms associated with online disinformation and manipulative practices. All come with the threat of substantial penalties for non-compliance. This is a move away from limited liability regimes that previously allowed platforms to thrive. Couple state actions with the measures taken by private actors, such as Facebook's development of a new Oversight Board,[2] and it is evident that there is a sea-change in the way content is regulated.

Platforms are a digital service that facilitate interactions between two or more distinct but interdependent sets of users who interact through the service via the Internet. They enable activities that are in most cases socially beneficial, and which may correspond to the exercise of fundamental rights and liberties: the freedom of expression, economic freedom, access to information, culture, and education, freedom of association, and political participation. Furthermore, they enable access to information and provide a forum for the exchange of ideas and opinions. However, prolific instances of platform manipulation by State actors, concerns about algorithmic dissemination of malicious content, and increasing evidence of hate speech have forced law and policymakers into rethinking the regulatory regime for platforms and for user-generated content. Platforms are not only of fundamental importance to free expression but are instrumental in the proper regulation of false information. They play a vital role in either the amplification or dissemination of falsehoods. Leaving the responsibility of policing content to self-regulating private parties shielded in a "safe harbour"[3] until they gain knowledge of problematic content is no longer seen as acceptable by governments. Furthermore, the principle of subsidiarity in supranational regimes ensures that the determination of the lawfulness of online content is largely left to national regulators. Unfortunately, this has opened the door to abuse and a "chilling effect" on free expression.

Until recently, democratic governments have stayed on the side-lines, avoiding any active role in content regulation. For example, the European Commission released a Communication on Disinformation arguing that the "primary obligation of state actors in relation to freedom of expression and media freedom is to refrain from interference and censorship and to ensure a favourable environment for inclusive and pluralistic debate".[4] However, the principle of subsidiarity acts as the legal justification for taking *proportional* measures to restrict the fundamental right to free expression. The common perception has been that national authorities are in a better position to strike the right balance between conflicting interests and the protection of the fundamental rights of the individual. However, with horizontal approaches falling out of favour, national authorities have either implemented or have proposed implementing a variety of vertical measures to end the days of self-regulation and general immunity for user-generated-content.

---

1 World Health Organization, 'Managing The COVID-19 Infodemic: Promoting Healthy Behaviours And Mitigating The Harm From Misinformation And Disinformation'. *Who.Int*, 2021, https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation.

2 An independent body that will hear appeals on content from Facebook users and advise the platform about its online speech policies, see 'Oversight Board | Independent Judgement. Transparency. Legitimacy.'. *Oversightboard.Com*, 2021, https://www.oversightboard.com/.

3 For example, the European Union's E-Commerce Directive, see 'E-Commerce Directive - Shaping Europe's Digital Future - European Commission'. *Shaping Europe'S Digital Future - European Commission*, 2021, https://ec.europa.eu/digital-single-market/en/e-commerce-directive.

4 'COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling Online Disinformation: A European Approach'. *Eur-Lex.Europa.Eu*, 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0236&from=RO. Accessed 11 Feb 2021.

Despite this change in approach, there has been a shift in thinking about how to shift the burden of content moderation onto private parties within frameworks for the effective protection of fundamental and human rights in a way that will:

- restrict or mitigate illegal or harmful behaviour,
- induce platforms to proactively prevent illegal or harmful behaviour from reaching their users,
- maintain platforms that are transparent and accountable for any content moderation,
- minimize any interference with fundamental rights to operate a business, and,
- avoid any 'collateral censorship'.

Content removal and/or moderation is a violation of fundamental and human rights; however, when the interference is "necessary in a democratic society", Courts will not require a formulation of "reasons why a comment was considered appropriate" and open the door for automatic takedown without assessment of the legality and appropriateness of the content. As it stands, there are four models of content regulation gaining traction around the world:

- The imposition of a "duty of care"
- Command-and-Criminality
- Enhanced Oversight
- Accountability-by-design

This report not only analyses forms of disinformation, but emerging models of content regulation. It critiques the move towards *proactive moderation* by private actors, while assessing threats to the freedom of expression posed by governments anxious to combat the threat of disinformation or to use this threat in order to justify their authoritarian approaches to regulating content. All of which could undermine democracy. Finally, based on the preceding analysis, we outline proposals on how to tackle the problem of fake news, disinformation, and online manipulation while effectively protecting fundamental rights.

Questions

1. What Do You Believe Are The Key Challenges Raised By Disinformation?

Disinformation to control the population

While there were only a few actors involved in social media manipulation around the time of the United Kingdom's referendum on continued membership of the European Union (Brexit) and the election of Donald Trump as the President of the United States, governments and political parties are now working with a wider range of actors including private firms, volunteer networks, and social media influencers to shape public opinion over social media. More sophisticated and innovative tools, including artificial intelligence and big data analytics, are being used to target, tailor, and refine messaging strategies. Governments and political parties are also increasingly relying on 'cyber troops' to manipulate online public opinion by either maliciously taking down legitimate content or accounts or to amplify false information.[5] Working with a wide range of actors - including private firms, volunteer networks, and social media influencers to shape public opinion over social media - governments and political parties are manipulating sections of the population that are most likely to be influenced by these messages.[6] By using these tactics, actors can discredit critics and de-legitimise legitimate media sources.[7] Anyone who challenges the narrative becomes a target for a high-volume online vilification.[8] Thus, by "tearing down the credibility of anyone questioning or critical of the government", dissenting voices are being silenced.[9] This can create a "chilling effect" on the freedom of expression, making others afraid to speak out against the government.[10]

Disinformation to disproportionately harm consumers and consumer confidence

Radu referred to the digital outbreak of disinformation during the COVID-19 pandemic as an "infodemic".[11] This has posed several challenges to the effective protection of consumers, particularly from various forms of fraud and cybercrime.[12] Online scammers have looked to take advantage of the crisis. These include offering fake COVID-19 tests, selling fake cures to the virus, or making bogus health claims.[13] EUROPOL has classified COVID-19 fraud into spurious websites, fake apps, fake investment opportunities, and money-muling[14] The European Commission has also said that many fraudulent websites have a COVID-19

5 Campbell-Smith, Ualan, and Samantha Bradshaw. *Global Cyber Troops Country Profile: India*. Oxford Internet Institute, University Of Oxford, 2021, p. 1, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-Profile.pdf. Accessed 15 Feb 2021.

6 Campbell-Smith, Ualan, and Samantha Bradshaw. *Global Cyber Troops Country Profile: India*. Oxford Internet Institute, University Of Oxford, 2021, p. 1, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-Profile.pdf. Accessed 15 Feb 2021.; Canadian Security Intelligence Service. *WHO SAID WHAT? The Security Challenges Of Modern Disinformation*. 2018, p. 7, https://www.canada.ca/content/dam/csis-scrs/documents/publications/disinformation_post-report_eng.pdf. Accessed 15 Feb 2021.

7 Id at Page 9.

8Id at Page 7.

9 Id at Page 82.

10 Id at Page 82.

11 Radu, Roxana. 'Fighting The 'Infodemic': Legal Responses To COVID-19 Disinformation'. *Social Media + Society*, vol 6, no. 3, 2020, p. 1. *SAGE Publications*, doi:10.1177/2056305120948190. Accessed 15 Feb 2021.

12 European Commission. *JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling COVID-19 Disinformation - Getting The Facts Right*. 2020, p. 3, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020JC0008&from=EN. Accessed 15 Feb 2021.

13 OECD. *Protecting Online Consumers During The COVID-19 Crisis*. 2020, pp. 2-3, https://read.oecd-ilibrary.org/view/?ref=130_130819-ay45n5rn74&title=Protecting-online-consumers-during-the-COVID-19-crisis. Accessed 15 Feb 2021.

14 'COVID-19: Fraud'. *Europol*, 2021, https://www.europol.europa.eu/covid-19/covid-19-fraud. Also see EUROPOL. *Catching The Virus Cybercrime, Disinformation And The COVID-19 Pandemic*. 2020, https://www.europol.europa.eu/publications-documents/catching-virus-cybercrime-disinformation-and-covid-19-pandemic. Accessed 15 Feb 2021.

related domain name that can be especially damaging when presented as a government or official website[15] The increased use of online communication tools due to social distancing has enabled a breeding ground for cybercrime[16] The Council of Europe has warned users about ransomware targeting mobile phones through apps that claim to supply truthful information about COVID-19, which are in fact fraudulent schemes aiming at extracting funds from consumers[17]

The manipulation of consumers, deceptive marketing techniques, and fraudulent activities may create profit when consumers are induced to buy unnecessary and dangerous products which are unsupported by scientific evidence.[18] One of the most notorious cases of disinformation was the claim by former US President Donald J. Trump that bleach injections could cure COVID-19.[19] According to the American Journal of Tropical Medicine and Hygiene,[20] these claims have cost lives.

Disinformation undermining democracy

Disinformation campaigns, emanating both from internal and external actors, are liable to adversely affect democratic processes and elections, threatening the foundations of democratic governments. Initially, disinformation was widely used with the intention to disrupt and distort elections and referenda, skewing political debate in favour or against specific candidates or undermining participation in elections.[21] Creators and propagators of fake news increasingly use social media, either manually or via the creation of automated accounts and political bots, to manipulate the information environment through an influx of fake news, with the intention to influence or polarise public opinion or promote scepticism towards electoral processes and institutions and, in the end, to undermine the integrity of democratic processes.[22] Moreover, social media and various messaging applications are used to spread disinformation via political advertisements.[23] In parallel, spreading fake news has been used by competitors in public debates, especially by candidates in elections, as a tool to influence electoral behaviour, i.e. to promote boycotts, to promote their candidacy, or attack the opposition.[24] On the other hand, "fact-checking" and politically influenced or affiliated "fake

15 J European Commission. JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling COVID-19 Disinformation - Getting The Facts Right. 2020, p. 16, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020JC0008&from=EN. Accessed 15 Feb 2021.

16 UNODC. 'CYBERCRIME AND COVID19: Risks and Responses'. 2020, p.1, https://www.unodc.org/documents/Advocacy-Section/UNODC_-_CYBERCRIME_AND_COVID19_-_Risks_and_Responses_v1.2_-_14-04-2020_-_CMLS-COVID19-CYBER1_-_UNCLASSIFIED_BRANDED.pdf. Accessed 15 Feb 2020.

17 Council of Europe. 'Cybercrime and COVID-19'. 27 March 2020, https://www.coe.int/en/web/cybercrime/-/cybercrime-and-covid-19. Accessed 15 Feb 2021.

18 European Commission. 'JOINT COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Tackling COVID-19 Disinformation - Getting The Facts Right'. 2020, p. 14, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020JC0008&from=EN. Accessed 15 Feb 2021.

19 BBC News. 'Coronavirus: Outcry after Trump Suggests Injecting Disinfectant Treatment.' BBC News, 24 Apr. 2020, www.bbc.com/news/world-us-canada-52407177. Accessed 15 Feb 2021.

20 Islam, Md Saiful, et al. 'COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis.' The American Journal of Tropical Medicine and Hygiene, vol. 103, no. 4, 2020, pp. 1621–29. Crossref, doi:10.4269/ajtmh.20-0812.

21 European Parliament. 'Policy Department for External Relations, Mapping Fake News and Disinformation in the Western Balkans and Identifying Ways to effectively Counter Them'. 2020, PAGES???, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/653621/EXPO_STU(2020)653621_EN.pdf. Accessed 15 Feb 2021.

22S. Bradshaw. ' Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project'. Oxford Internet Institute. 2018, p. 5. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf. Accessed 15 Feb 2021.

23 Ibid, p. 13.

24 European Parliament. '*Policy Department for External Relations, Mapping Fake News and Disinformation in the Western Balkans and Identifying Ways to effectively Counter Them'.* 2020, pp. 30-45, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/653621/EXPO_STU(2020)653621_EN.pdf. Accessed 15 Feb 2021.

news observatories" have been used to influence political debate by regularly flagging statements of the opposition as fake news, to spread government-led propaganda or to spread disinformation.[25]

Disinformation causing societal harm

Disinformation may also cause societal harms. It undermines access to trustworthy information and affects the credibility of both government information and mainstream journalism. Social media has demonstrated to be particularly useful in causing those societal harms. Through social media, many people can be simultaneously reached with personalised messages and micro-targeting advertisements. It has been demonstrated that social media manipulation campaigns have been carried out in 48 countries with a powerful effect to subvert elections and undermine trust in democratic institutions.[26] In each country, there is often at least one political party or government agency using social media to manipulate public opinion domestically. In the Philippines, the social news network "Rappler" has documented hundreds of websites and millions of social media accounts and groups that methodically and consistently spread disinformation.[27] Furthermore, disinformation has incited violence, for example in India five people were publicly lynched following the dissemination of fake rumours on WhatsApp about "outsiders abducting children"[28]Similarly, in Mexico, 100 people burned two men alive due to the circulation of false rumours on WhatsApp about a "plague of child kidnappers.[29] Moreover, computational propaganda involves not only social media account automation and online commentary teams but also includes paid advertisements and search engine optimisation on a widening array of Internet.[30] Thus, social media manipulation damages society, as well as increasingly constituting a profitable business.

Disinformation amplifying discriminatory practices

Disinformation thrives in societies characterised by ethnic, religious, gender and socio-political diversity. False information can incite and amplify hatred. It represents a means to discriminate, to create social disorder, and political destabilisation. For example, in North Macedonia, disinformation campaigns about alleged NATO and EU support for the implementation of the "Greater Albania" idea was used to incite tensions between ethnic Albanians and Macedonians.[31] Disinformation may also lead to discrimination against minorities and other vulnerable groups. For example, in the Western Balkans, false information described migrant and refugee communities as responsible for the spread of COVID-19.[32]

False narratives can also cause radicalisation. To promote anti-refugee sentiment and Islamophobia, fake news can incite xenophobic and racist behaviours, while legitimising the discriminatory practices of governments. Moreover, disinformation can also be used as "proof" that supports discriminatory theories used to implement exclusionary policies. The Minister of Justice for Greece referred to studies that children raised by parents of different genders developed fewer psychological problems than ones raised in a same-

25 Interesting example constitutes the 'Fake News Observatory' of the now-leading party in Greece. See https://thepressproject.gr/to-paratiritirio-fake-news-tis-nd-xanachtypa-me-psemata-para-tin-ypanachorisi-ston-tromonomo-stin-techni/ (in Greek).

26 S. Bradshaw. 'Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project'. Oxford Internet Institute. 2018, p. 5. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf. Accessed 15 Feb 2021.

27 Ibid, p. 81.

28Dixit, Pranav. 'How WhatsApp Destroyed A Village.' BuzzFeed News, 7 Nov. 2018, www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india. Accessed 15 Feb 2021.

29 Funke, Daniel. 'Misinformation Is Inciting Violence around the World. And Tech Platforms Don't Seem to Have a Plan to Stop It.' Poynter, 4 Apr. 2019, www.poynter.org/fact-checking/2019/misinformation-is-inciting-violence-around-the-world-and-tech-platforms-dont-have-a-plan-to-stop-it. Accessed 15 Feb 2021.

30 S. Bradshaw. 'Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project'. Oxford Internet Institute. 2018, p. 5. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf. Accessed 15 Feb 2021.

31 European Parliament. 'Policy Department for External Relations, Mapping Fake News and Disinformation in the Western Balkans and Identifying Ways to effectively Counter Them'. 2020, p. 30, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/653621/EXPO_STU(2020)653621_EN.pdf. Accessed 15 Feb 2021.

32 Ibid, p. 43.

sex family.[33] Furthermore, the spread of disinformation also disproportionately effects women, especially those involved in politics.[34] Research has shown that there is a correlation between the spread of disinformation relating to the roles, campaigns, beliefs, and actions of women in politics, and the harassment, abuse, and vitriol directed back at them.[35] This has significant implications for women and girls as for them, the "chilling effect" of disinformation is likely to result in a reduction in their participation.[36]

Disinformation causing individual harms

While disinformation affects national security, democracy, consumer protection and society, it also directly affects individuals. It can result in mistrust in the democratic process and in the national media, which can lead to disengagement from civil and political life.[37] False news, because of its nature, has the consequence of creating confusion and provoking fear, anxiety, and fatigue. It may undermine mental integrity and self-development since disinformation is damaging to the search for the truth. Finally, disinformation may cause real practical difficulties in citizens' everyday life, as proven during the COVID-19 pandemic.

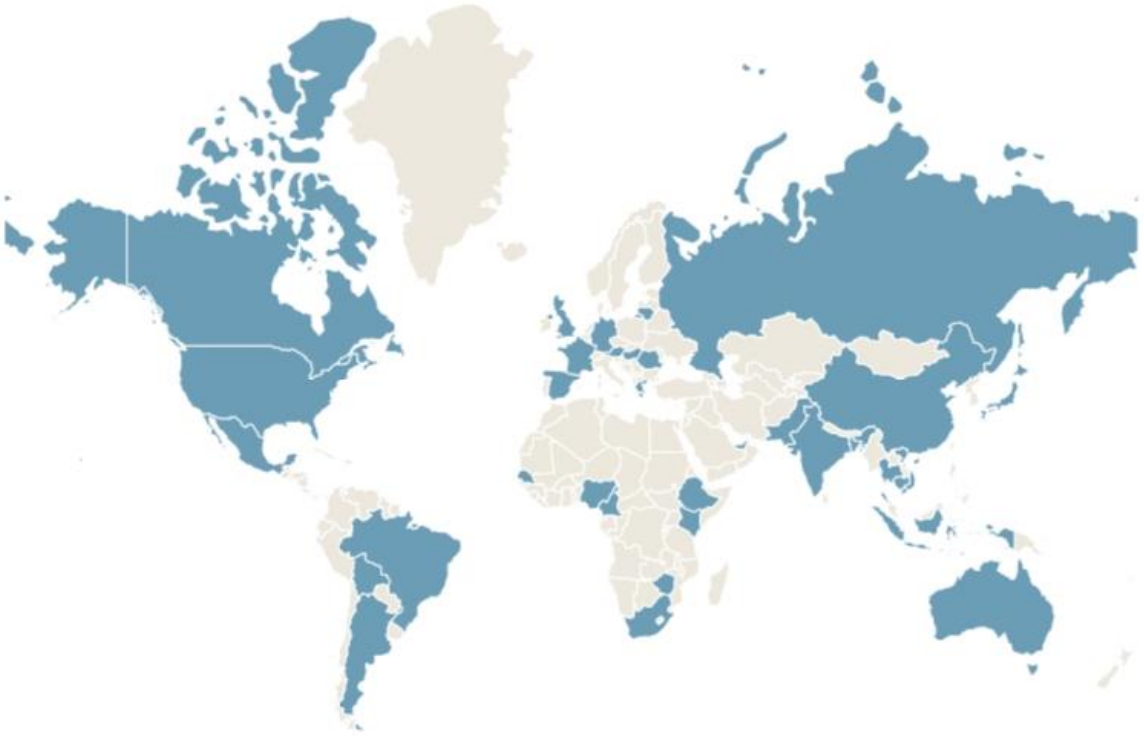
2(a). *What Legislative, Administrative, Policy, Regulatory Or Other Measures Have Governments Taken To Counter Disinformation Online And Offline?*

General measures taken by Governments to combat fake news

In the last few years, combating fake news and disinformation has been at the forefront of lawmakers' agendas. Often the adopted measures are conducted by new government agencies or through existing organisations. Their activities aim to issue counter-narratives or create reporting, flagging, and fact-checking portals to support citizen awareness and engagement. However, it is not only democracies that have responded to the pressing issues posed by disinformation, but also authoritarian regimes. In many cases, the task forces responsible for combating disinformation are the ones who incentivise further censorship, usually using media laws, increased surveillance capabilities, computational propaganda campaigns, and Internet blocking or filtering. Therefore, the actual aim is, not to limit false news, but to restrict freedom of expression and shape online public discourse in a manner that is favourable to the ruling party.[38]

This section will outline measures taken to combat disinformation by supranational bodies and Governments in Europe, the Americas, the Asia-Pacific region, and Africa, before turning to the impact of these measures on human rights. Lastly, it will consider actions taken by Governments and human rights mechanisms to address the negative impact on rights.

33Αυγή Newsroom. 'Κώστας Τσιάρας / Δήλωση - Προσβολή Για Μονογονεϊκές Οικογένειες Και Ομόφυλα Ζευγάρια.' Αυγή, 2 Feb. 2021, www.avgi.gr/koinonia/378354_dilosi-prosboli-gia-monogoneikes-oikogeneies-kai-omofyla-zeygaria. Accessed 15 Feb 2021.
34OHCHR. 'The Impact of Online Violence on Women Human Rights Defenders and Women's Organisations.' OHCHR, 21 June 2018, www.ohchr.org/EN/HRBodies/HRC/Pages/NewsDetail.aspx?NewsID=23238&LangID=E. Accessed 15 Feb 2021.
35 Barker, Kim and Jurasz, Olga. 'Gendered Misinformation & Online Violence Against Women in Politics: Capturing legal responsibility'. Co-Inform. https://coinform.eu/gendered-misinformation-online-violence-against-women-in-politics-capturing-legal-responsibility/. Accessed 15 Feb 2021.
36 Ibid.
37 S. Bradshaw. 'Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project'. Oxford Internet Institute. 2018, p. 3. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf. Accessed 15 Feb 2021.
38 Ibid, p. 6.



This map highlights the nations we will now analyse

I. Europe

The Council of Europe (CoE) does not have any concrete legal framework to counter disinformation online and offline. However, the CoE has a rich body of internet governance and guiding principles which should apply both online and offline.[39]

European Union (EU) In the EU, countering disinformation is a significant policy aim, to the extent that it is considered one of three main pillars in the European Democracy Action Plan.[40] The EU approach to tackling online disinformation consists of a series of measures to promote a more transparent, trustworthy, and accountable online ecosystem, to ensure resilient election processes, to foster education and media literacy, to support quality journalism, and to counter internal and external disinformation threats through strategic communication.[41]

39 See for example, Council of Europe. 'Recommendation CM/Rec(2016)5 of the Committee of Ministers to member States on Internet freedom'. 13 April 2016; Council of Europe. 'Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a Guide to human rights for Internet users'. 16 April 2014; Council of Europe. 'Recommendation CM/Rec(2018) of the Committee of Ministers to Member States on the roles and responsibilities of internet intermediaries'. 7 March 2018; Council of Europe. 'Recommendation CM/Rec(2012)4 of the Committee of Ministers to Member States on the protection of human rights with regard to social networking services'. 4 April 2012.

40 European Commission. 'European Democracy Action Plan: making EU democracies stronger'. 3 December 2020, https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2250. Accessed 14 Feb 2021.

41 European Commission. 'Tackling online disinformation: a European Approach". 24 April 2018, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236. Accessed 13 Feb 2021.

The EU Action Plan on Disinformation[42] highlights cooperation to counter disinformation, not only between Member States, but also with platforms and industry. This is the first time that a Code of Practice on Disinformation, which includes self-regulatory standards to fight disinformation and fake news, has been agreed to by industry. It has been signed by online platforms and the advertising industry on a voluntary basis and is checked regularly by the European Commission.[43] During the COVID-19 pandemic, joint efforts in tackling disinformation have been emphasized to protect public health and consumers' rights, to raise citizen awareness, and to ensure common values and democracy.[44]

The East StratCom Task Force was created within the European External Action Service, with the purpose of addressing ongoing disinformation campaigns in the Baltics by the Russian Federation. The Task Force has set up an EU v Disinformation webpage, which functions as an EU-wide rapid alert system and is meant to "facilitate the sharing of insights related to disinformation campaigns and coordinate responses".[45] The system is based on open-source information and draws from "academia, fact-checkers, online platforms and international partners" expertise.[46]

The European Digital Media Observatory, which is funded to create a European hub for fact-checkers, academics and other relevant stakeholders to collaborate with each other, provides support for European policy makers in actions against disinformation.[47] Moreover, the EU is also funding a joint EU-wide network for fact-checkers. In total, €2.5 million has been budgeted for this purpose.[48]

On 15 December 2020, the European Commission published its proposal for the Digital Services Act, which is considered to contribute to fighting disinformation. It will aim to ensure that platforms are more accountable and responsible for the systemic risks they pose concerning online disinformation. It introduces new rules on how platforms moderate content, on advertising, algorithmic processes and risk mitigation. Furthermore, it sets out a co-regulatory framework where service providers can work under codes of conduct to address negative impacts emanating from manipulative activities, including online disinformation.[49]

In **Germany**, the Constitution does not guarantee a general right to freedom of expression which may be interpreted as including the right to fabricate and send false statements or deceptive expressions; instead, only the expression of opinions is constitutionally protected,[50] while, according to the German constitutional court, "incorrect information is not an interest that merits protection".[51] From a regulatory perspective, according to the Law for the Media, electronic information and communication services (TeleMedia), including social media platforms, providing journalistic content are responsible of abiding by the "recognized journalistic standards", including the duty to verify the veracity of information.[52] However, the spreading of disinformation remains generally unsanctioned: the Press Code (Pressekodex) applies only to persons and entities that have voluntarily adhered to it, while the Press Council (Deutscher Presserat) has the power only to issue public reprimands.[53] Moreover, according to the Telemedia Act, host providers, including social media, are not liable for false or inaccurate information published on their platforms

42 Ibid. PAGE?

43 European Commission. 'Code of Practice on Disinformation'. https://ec.europa.eu/digital-single-market/en/code-practice-disinformation. Accessed 14 Feb 2021. PAGE??

44 European Commission. 'Tackling COVID-19 disinformation - Getting the facts right'. 10 June 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020JC0008. Accessed 14 Feb 2021. PAGE?

45 EU v Disinformation website. https://euvsdisinfo.eu/. Accessed 15 Feb 2021.

46 Ibid.

47 European Commission. 'Tackling online disinformation'. https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation. Accessed 14 Febc2021.

48 Library of Congress. 'Government Responses to Disinformation on Social Media Platforms: European Union' . https://www.loc.gov/law/help/social-media-disinformation/eu.php. Accessed 15 Feb 2021.

49 European Commission. 'Digital Services Act – Questions and Answers'. 15 December 2020, https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348#1. Accessed 14 Feb 2021.

50 Article 5, German Constitution.

51 BVerfGE 85, 1 - Bayer-Aktionäre.

52 Staatsvertrag für Rundfunk und Telemedien [Rundfunkstaatsvertrag] [RStV], Aug. 31, 1991, as amended, Art. 54, para. 2.

53 Presserat, Publizistische Grundsätze [Pressekodex] (2017), Complaints Procedure, § 12, para. 5, in conjunction with § 16.

insofar as they don't have actual knowledge of the existence of infringing content; it is only upon notification that they are obliged to remove content to escape liability.[54] The unsatisfactory response of online platform operators to remove illicit content, including fake news and hate speech, triggered the intervention of the German legislator: the 2017 Germany's Network Enforcement Act (NetzDG)[55] uses a strict command-and-control approach, requiring platforms that have two million or more registered users in Germany to remove unlawful content within seven days after flagging and within twenty-four hours where the content is "manifestly unlawful". Should they fail to do so, providers face severe fines up to €50 million, imposed by the Ministry of Justice upon a binding court decision. Taking into consideration the explanatory memorandum of the Act, fake news constitutes illicit content, among other cases specifically described in German criminal laws, when they promote hatred, abuse, defamation, propaganda, when they could lead to a breach of the peace by misleading authorities into thinking a crime has been committed, and when they constitute public incitement to crime. To comply with the Network Enforcement Act, operators of social media platforms must offer to users easy-to-use, permanent and transparent flagging mechanisms; decisions upon complaints should be communicated to the complainant and the affected individuals. Moreover, platforms receiving more than one hundred complaints per year must publish biannual reports in German about the mechanisms employed for handling of complaints, both in the Federal Gazette and on the homepage of the social media network.

In **France,** art. 27 of the Law on Freedom of the Press criminalizes the mala fide publication, dissemination, or reproduction of news wholly or partly falsified insofar as they are liable for disrupting public peace; the fine for infringing this provision can reach €45,000.[56] France has also undertaken measures against disinformation during the pre-election period: the dissemination of fake news or other "fraudulent schemes" that have the potential of affecting electoral results are punishable by up to one year in jail and a fine of up to €15,000 under art. 97 of the French Electoral Code.[57] Moreover, new legislation[58] enacted in 2018 aims at countering large-scale dissemination of falsified information through online platforms, requiring platform operators to increase transparency and take measures to stop disinformation campaigns during the three months preceding national elections. Specifically, Internet platform operators are required to provide users with 'honest, clear and transparent information' about the identity of anyone who paid to promote information related to a debate of national interest" as well as about the use of personal data in the context of promoting election-related content; disclosure of the amounts paid for the promotion of content related to the pre-electoral public debate exceeding a certain threshold is also mandated.[59] Furthermore, judges have the discretion to order 'any proportional and necessary measure' to stop the 'deliberate, artificial or automatic and massive' dissemination of false or misleading information on the

54 Telemediengesetz [TMG], Feb. 26, 2007, BGBl. I at 179, as amended, § 10.

55 Telemediengesetz [TMG] [Telemedia Act]. Feb. 26, 2007, BGBl. http://www.gesetze-im-internet.de/tmg/TMG.pdf, archived at http://perma.cc/3YJK-9N48. Accessed 15 Feb 2021. Unofficial English translation available at https://www.huntonprivacyblog.com/wp-content/uploads/sites/28/2016/02/Telemedia_Act__TMA_.pdf (English version not updated). Accessed 15 Feb 2021.

56 Law of 29 July 1881 on Freedom of the Press.
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006070722&dateTexte=vig. Accessed 13 Feb 2021.

57 Electoral Code. https://www.legifrance.gouv.fr/affichCodeArticle.do?
cidTexte=LEGITEXT000006070239&idArticle=LEGIARTI000006353232. Accessed 14 Feb 2021.

58 Organic Law No. 2018-1201 of 22 December 2018 Regarding the Fight Against Information Manipulation.
https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=3EA914DFE69980E3FBB01324A666B5D1.tplgfr22s_1?cidTexte=JORFTEXT000037847556&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000037847553. Accessed 13 February 2021; Law No. 2018-1202 of 22 December 2018 Regarding the Fight Against Information Manipulation.
https://www.legifrance.gouv.fr/affichTexte.
do;jsessionid=3EA914DFE69980E3FBB01324A666B5D1.tplgfr22s_1?cidTexte=JORFTEXT000037847559&categorieLien=id. Accessed 13 Feb 2021.

59 Organic Law No. 2018-1201 of 22 December 2018 Regarding the Fight Against Information Manipulation.
https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=3EA914DFE69980E3FBB01324A666B5D1.tplgfr22s_1?cidTexte=JORFTEXT000037847556&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000037847553. Accessed 13 February 2021.

Internet that could disturb public order or affect the validity of the elections, after request from the public prosecutor, a candidate, a political group or party, or any person, which is adjudicated within 48 hours.[60]

Platform operators are additionally required to introduce an easily accessible system enabling users to flag disinformation related to the elections and are encouraged to introduce further measures to increase the transparency of their algorithms, promote content from companies and press agencies and audio-visual communication services, fight against accounts that massively propagate false information, inform users of the nature, origin and distribution methods of the content and promote media literacy;[61] operators are also required to provide, on a yearly basis, a report to the Superior Council on Audio-visual Services (Conseil Supérieur de l'audiovisuel) regarding the implementation of the measures.[62]

Finally, under the recent French law against online hate speech, Internet platforms are required to delete, with 24 hours upon notification, content deemed 'manifestly unlawful' on grounds of race, religion, sex, sexual orientation or disability; failure to comply may result in fines up to 4% of the turnover of the previous fiscal year. Further, platforms would have to implement a system allowing: (i) in the event of the removal of content, the user who initiated the publication of the removed content to contest this removal; and (ii) in the event of non-removal of signed content, the author of the report, to challenge the maintenance of the content.[63]

In **Russia**, two laws promulgated in 2019 criminalize the malicious spread of 'socially significant' fake news, defined as false information distributed as truthful that entails a threat towards people's lives, health, or property, public order or public security, transportation and social infrastructure, credit institutions, lines of communications, industry, and energy enterprises.[64] Specifically, the 'Law on Amending Article 15-3 of the Information Law' entrusts to the 'Federal Service for Supervision of Communications, Information Technology and Mass Media' (Roskomnadzor) the power to request the editorial body of an on-line publication to remove fake news; should the latter fail to remove immediately the said material from the website, the Roskomnadzor is entitled to take measures to limit access thereto and mandate Internet Service Providers to immediately block access to websites disseminating the fake information in question; the Law on Amending the Code of Administrative Violations establishes pecuniary sanctions for spreading fake news.[65] Moreover, the owners of online news aggregators, who should mandatorily be Russian natural or legal persons, are liable for the verification of the truthfulness and reliability of socially significant disseminated information.[66]

In **Greece**, under article 191 of the Criminal Code the dissemination or spreading via the Internet of false news 'causing fear' is a criminal offence, punishable with up to three years of imprisonment; negligent spreading of fake news also constitutes a criminal offence.[67]

Malta's Criminal Code criminalizes the malicious spreading of false news that is likely to alarm public opinion or disturb public good order or public peace or to create a commotion among the public or among

60Ibid.
61 Ibid.
62 Ibid.
63 LAW n ° 2020-766 of June 24, 2020 aimed at combating hateful content on the internet. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000042031970/. Accessed 12 Feb 2021.
64 Federal Law No. 31-FZ of March 18, 2019, on Amending Article 15-3 of the Federal Law on Information, Information Technologies and Protection of Information. http://publication.pravo.gov.ru/Document/View/0001201903180031 (in Russian). Accessed 15 Feb 2021; Federal Law No. 27-FZ of March 18, 2019 on Amending the Code of Administrative Violations. http://publication.pravo.gov.ru/Document/View/0001201903180021? index=1&rangeSize=1 (in Russian). Accessed 15 Feb 2021.
65 Library of Congress. 'Initiatives to Counter Fake News: Russia', Index of Initiatives to Counter Fake News. Updated 30 December 2020, https://www.loc.gov/law/help/fake-news/russia.php. Accessed 15 Feb 2021.
66 Federal Law on Information. 'Information Technologies and Protection of Information N149-FZ'. July 27 2006, http://pravo.gov.ru/proxy/ips/? docbody&nd=102108264. Accessed 15 Feb 2021.
67 Law 4619/2019 - Government Gazette 95 / A / 11-6-2019 - Ratification of the Penal Code. https://www.e-nomothesia.gr/kat-kodikes-nomothesias/nomos-4619-2019-phek-95a-11-6-2019.html. Accessed 15 Feb 2021.

certain classes of the public; infringements of this provision are punishable with imprisonment from one to three months.[68]

In **Lithuania,** the Law on the Provision of Information to the Public prohibits the intentional dissemination of false information by public information producers, disseminators, participants therein, journalists and related institutions, establishing rules for the liability thereof.[69]

Austria criminalizes dissemination of false news during an election or referendum.[70]

The **Spanish** government adopted last November a ministerial order to combat disinformation, that includes the creation of a permanent committee (Comité de la Verdad) in charge of monitoring online disinformation campaigns and implement policy measures.[71] This new protocol has been backed up by the European Commission as part of Spain's participation in the European Union's Action Plan Against Disinformation.[72]

In **Hungary,** the 'Anti-Coronavirus Act' criminalizes the dissemination of 'false' or 'distorted' COVID-19 information, punishable with imprisonment up to 5 years.[73]

In **Romania,** under a Presidential Decree relating to the state of emergency due to the Coronavirus pandemic, public institutions and authorities are authorised to 'undertake the necessary measures in order to correctly and objectively inform the population' in case of dissemination of fake-news in mass-media and on-line in relation to COVID-19. Specifically, hosting and content service providers are required to immediately interrupt the transmission and remove such content from its source or block access to that content and inform the users, upon decision of the National Authority for Regulation in Communication. Furthermore, when providers do not fall within Romanian jurisdiction, they are required to immediately block access to infringing content to users located within the territory of Romania.[74]

The **United Kingdom** does not yet have comprehensive legislation countering fake news and disinformation.[75] The Online Harms Bill, pending a vote in early 2021, aims at clarifying the responsibilities of online companies in relation to the safety of Internet users, establishing a new statutory 'duty of care' thereof to take reasonable steps to counter illegal and harmful content or activity, including the dissemination of fake news.[76] According to the Online Harms White Paper, as made available for public consultation, companies offering content online will need to take proportionate and proactive measures to help users understand the reliability of the information they are receiving, to minimise the spread of misleading and harmful disinformation and to increase the accessibility of trustworthy and varied news

68 Malta, Art. 82 of Criminal Code. https://justice.gov.mt/en/pcac/Documents/Criminal%20code.pdf. Accessed 15 Feb 2021.
69 Republic of Lithuania, Law on Provision of Information to the Public, 2 July 1996 No. I-1418. Official translation in English available at:
https://www.legislationline.org/download/id/5542/file/Lithaunia_law_provision_information_public_am2006_en.pdf. Accessed 15 Feb 2021.
70 Austria, Section 264 of Austria's Criminal Code.
71 El País. 'Spain to monitor online fake news and give a 'political response' to disinformation campaigns'. 9 November 2020, https://english.elpais.com/politics/2020-11-09/spain-to-monitor-online-fake-news-and-give-a-political-response-to-disinformation-campaigns.html?rel=listapoyo. Accessed 15 Feb 2021.
72 El País. EU Commission backs Spain's protocol against disinformation campaigns. 10 November 2020, https://english.elpais.com/politics/2020-11-10/eu-commission-backs-spains-protocol-against-disinformation-campaigns.html. Accessed 15 Feb 2021.
73 Radu, Roxana. 'Fighting The 'Infodemic': Legal Responses To COVID-19 Disinformation'. Social Media + Society, vol 6, no. 3, 2020, p.2. SAGE Publications, doi:10.1177/2056305120948190. Accessed 15 Feb 2021. ; Access Now. 'Fighting disinformation and defending free expression during covid-19: recommendations for States'. 2020, p. 13, .
74 Romania, Decree on the extension of the state of emergency in the territory of Romania, Official Journal of Romania, Part I, No. 311/14.04.2020. Available in English at https://rm.coe.int/16809e375e. Accessed 15 Feb 2021.
75 Library of the Congress. 'Government Responses to Disinformation on Social Media Platforms: United Kingdom'. https://www.loc.gov/law/help/social-media-disinformation/uk.php. Accessed 15 Feb 2021.
76 United Kingdom. Consultation outcome - Online Harms White Paper. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper. Accessed 15 Feb 2021.

content.[77] Compliance with the duty of care is proposed to be monitored by an independent Regulator, who would have the power to bring legal action against platform operators and impose heavy fines in case of breach of this duty.[78] Moreover, the regulator is expected to produce a 'code of best practice', including mandatory rules for companies falling in scope thereof, which may include, indicatively, the use of fact-checking services, the promotion of authoritative news sources, improving the transparency of political advertising etc.[79]

II. The Americas

Argentina: With regards to the transparency of political ads, a recent amendment to Argentinian law on the financing of political parties has included specific measures aimed at reinforcing the transparency and accountability of online political advertising by requiring those ads to fully disclose the purchaser's identity as well as the registration of political parties' social media accounts[80]. Furthermore, Argentina is considering the creation of a Comisión de Verificación de Noticias Falsas (CVNF) (Commission for the Verification of Fake News) under the framework of the Camara Nacional Electoral (CNE) that would oversee the detection, recognition and prevention of online fake news during electoral campaigns. This Commission would carry out the review of publications to verify their authenticity, excluding those focused on ideological positions. The CVNF would then report to the CNE, which would have the power to release binding orders to internet service providers.[81]

Bolivia: In March 2020, the Bolivian government approved a decree that includes a provision saying that 'individuals who incite non-compliance with this decree or misinform or cause uncertainty to the population will be subject to criminal charges for crimes against public health'[82]. The decree punishes violations with jail time up to 10 years.[83]

Brazil: The Brazilian Senate passed in June 2020 the 'The Brazilian Internet Freedom, Responsibility and Transparency Act',[84] which criminalizes with prison penalties actions such as creating or sharing content that allegedly poses a serious risk to 'social peace or to the economic order'. The Bill also requires internet service providers to collect user's ID and cell phone numbers to open email or messaging accounts as well as tracking the chain of communications for at least four months.[85] The Bill still needs approval from congress. Furthermore, since the last municipal elections, in October 2020, the Superior Electoral Court (Tribunal Superior Electoral in Portuguese) created the position of Digital Coordinator to Combat Disinformation. The position was given to a technician with extensive experience on internet manipulation. This measure helped to develop technical partnerships with tech companies, including some of the tech

77 Ibid.
78 Ibid.
79 Ibid.
80 Library of Congress. 'Government Responses to Disinformation on Social Media Platforms: Comparative Summary'. https://www.loc.gov/law/help/social-media-disinformation/compsum.php. Accessed 15 Feb 2021.
81 Library of Congress. 'Initiatives to Counter Fake News in Selected Countries'. 2019, pp. 4-6, https://www.researchgate.net/profile/Jenny_Gesley/publication/345361074_Initiatives_to_Counter_Fake_News_in_Selected_Countries_Argentina_Brazil_Canada_China_Egypt_France_Germany_Israel_Japan_Kenya_Malaysia_Nicaragua_Russia_Sweden_United_Kingdom/links/5fbd274f458515b797692f2c/Initiatives-to-Counter-Fake-News-in-Selected-Countries-Argentina-Brazil-Canada-China-Egypt-France-Germany-Israel-Japan-Kenya-Malaysia-Nicaragua-Russia-Sweden-United-Kingdom.pdf. Accessed 15 Feb 2021.
82 Committee to Protect Journalists. 'Bolivia enacts decree criminalizing 'disinformation' on COVID-19 outbreak'. April 2020, p. 9, https://cpj.org/2020/04/bolivia-enacts-decree-criminalizing-disinformation/. Accessed 15 Feb 2021.
83 Human Rights Watch. 'Bolivia: COVID-19 Decree Threatens Free Expression'. April 7 2020, https://www.hrw.org/news/2020/04/07/bolivia-covid-19-decree-threatens-free-expression. Accessed 15 Feb 2021.
84 Reuters. 'Brazil Senate approves bill on fake news, lower house to vote next'. July 1 2020, https://www.reuters.com/article/us-brazil-politics-fake-news-idUSKBN2413T1. Accessed 15 Feb 2021.
85 Human Rights Watch, Brazil: Reject 'Fake News' Bill, June 24 2020, available at https://www.hrw.org/news/2020/06/24/brazil-reject-fake-news-bill

giants, to prevent the spread of disinformation. The reactions of academia to this measure have been incredibly positive, even though it is still necessary to see what will happen in a presidential election.[86]

Canada: While Canada doesn't have a law prohibiting the dissemination of false information (unless it is defamatory, in which case it is covered by libel laws), the government has launched several initiatives particularly in the field of election ads. The government passed in 2017 an omnibus bill that amended the Canada Elections Act which included 'a provision that makes it an offence to make false statements about a candidate for the purpose of influencing the outcome of an election.'[87] In January 2019, the government announced a series of measures aiming at further shoring up Canada's electoral system from foreign interference, and enhancing Canada's readiness to defend the democratic process from cyber threats and disinformation.[88] One of the initiatives included the creation of a Critical Election Incident Public Protocol,[89] that would monitor disinformation attempts and notify other agencies and the public.[90] More recently, the government announced in November 2020 the launch of Canada's Digital Act, which inter alia aims at defending freedom of expression and protecting against online threats and disinformation.[91]

Mexico: While Mexico does not have specific legislation on the topic of disinformation, the government is countering this phenomenon with the use of official channels of communication. For example, Mexico's National Institute of Elections responds to disinformation through information on social media platforms and has signed collaborative agreements with Facebook and Google with regards to the 2018 elections. In addition, the newswire service Notimex run by the Mexican President's staff launched the initiative 'Verificado', aiming at fact-checking news on social media and traditional media.[92]

The United States: The First Amendment of the Constitution has so far prevented the adoption of any laws curtailing freedom of speech or free press. The jurisprudence developed by the Courts under this amendment has not been receptive to regulating speech based on its content, particularly when the speech is political. Any restriction is examined under a strict scrutiny test which almost always results in the regulation being struck down.[93] The U.S. Supreme Court has only recognised certain forms of speech as incapable of availing themselves of the protection of the first amendment: fraud, obscenities, defamation, and incitement.[94] However, and in view of Russia's meddling in the 2016 elections, some initiatives have been created to tackle disinformation. In October 2017, the Congress announced a new billed called the Honest Ads Act that would require companies like Google and Facebook to keep track of political ads and fully disclose them.[95] Furthermore, this act would also require these companies to disclose the targets of these ads as well as the information on the buyer and the rates charged.[96] Notably, in November 2017 representatives of Facebook, Google and Twitter testified before the Senate Judiciary Committee about their role in Russian interference with the elections through online platforms.[97] More recently in November

86 ITS, TSE cria ação contra Fake News, 13 October 2020, available at https://itsrio.org/pt/artigos/tse-cria-acao-contra-fake-news/
87 'Initiatives To Counter Fake News'. Loc.Gov, 2020, https://www.loc.gov/law/help/fake-news/canada.php. Accessed 11 Feb 2021.
88 Rachel Aiello, Feds Unveil Plan to Tackle Fake News, Interference in 2019 Election, The Star (Feb. 8, 2018), https://www.thestar.com/news/canada/2018/02/08/trudeau-to-facebook-fix-your-fake-news-problem-or-else.html
89 https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html
90 Daniel Funke, A Guide to Anti-disinformation Actions around the World, Poynter, available at https://www.poynter.org/ifcn/anti-disinformation-actions/
91 Government of Canada, Canada's Digital Charter: Trust in a digital world, available at https://www.ic.gc.ca/eic/site/062.nsf/eng/h_00108.html
92 https://notimex.mx/es/verificado/
93 Nuñez, Fernando. Disinformation Legislation and Freedom of Expression, 10 U.C. Irvine L. Rev. 783 (2020), p. 789 available at https://scholarship.law.uci.edu/ucilr/vol10/iss2/10
94 United States Supreme Court, United States v. Alvarez, 567 U.S. 709, 717 (2012).
95 U.S. Congress, S.1989 'Honest Aids Act', 115th Congress, available at https://www.congress.gov/bill/115th-congress/senate-bill/1989, accessed 15/2/2021.
96 U.S. Congress, S.1989 'Honest Aids Act', 115th Congress, available at https://www.congress.gov/bill/115th-congress/senate-bill/1989, accessed 15/2/2021.
97 CNN, Facebook, Twitter, Google defend their role in election, November 1 2017, available at https://money.cnn.com/2017/10/31/media/facebook-twitter-google-congress/index.html

2020, the representatives of Facebook and Twitter testified about their platforms and the moderation of disinformation in the context of the 2020 elections.[98]

III. Asia-Pacific

Australia: In 2018, an Electoral Integrity Assurance Task Force was set up to address risks to the integrity of the electoral system—particularly in relation to cyber interference. The *National Security Legislation Amendment (Espionage and Foreign Interference) Act 2018* inserted new foreign interference offences into the Commonwealth *Criminal Code*.[99] The elements of these foreign interference offences could arguably be applied to persons who 'weaponize' fake news in certain circumstances.[100] The Australian Election Commission (AEC) commenced an advertising campaign, *Stop and consider* on social media platforms (such as Facebook, Twitter and Instagram) to encourage voters to 'carefully check the source of electoral communication they see or hear' during the 2019 federal election campaign.[101] The AEC is reported to have sought to establish protocols for social media companies to address advertising on their platforms that contravene Australia's electoral laws (such as those relating to authorisation). In May 2017, the Senate established the Select Committee on the Future of Public Interest Journalism. The impact that digital disruption has had on the business model of traditional media was a key focus of the inquiry. It also considered the effect that search engines and social media have on public interest journalism with the spread of fake news. The committee said it was 'struck by the number of journalist jobs that have been lost in the last few years' due to industry restructuring.[102] It received evidence that such changes were, in some cases, leading to a decline in quality journalism and continuing to erode the public's trust in media. The Senate committee's report also made several recommendations regarding supporting public interest journalism, including adequate levels of funding for the national broadcasting sector and an audit of current laws that impact on journalism.[103] The Senate committee specifically referred to fake news in relation to education, proposing that the Commonwealth work with states and territories to see if the national curriculum can be strengthened—not only to enhance student awareness of fake news, but to improve digital media literacy skills more generally.[104]

Bangladesh: In October 2018, Bangladesh adopted a controversial law called the Digital Security Act, which is the main law the government now uses to deal with fake news on the web and social media.[105] The act stipulates that the publishing or sending of offensive, false or fear inducing data-information and the publication of information with the intent of tarnishing the image of the nation or spreading confusion with

98 The New York Times, Zuckerberg and Dorsey Face Harsh Questioning From Lawmakers, available at https://www.nytimes.com/live/2020/11/17/technology/twitter-facebook-hearings

99 Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, <

100Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews, accessed 15/2/2021

101Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews, accessed 15/2/2021

102Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews, accessed 15/2/2021

103Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews, accessed 15/2/2021

104Buckmaster, Luke, and Wils Tyson, 'Responding to fake news', Parliament of Australia, https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews, accessed 15/2/2021

105 Library of Congress, 'Freedom of Expression during COVID-19', Index of Legal Reports, September 2020, https://www.loc.gov/law/help/covid-19-freedom-of-expression/bangladesh.php#_ftn7, accessed 15/2/2021.

the intention to do so will be an offence under the act, which will be penalised with imprisonment or a fine.[106]

Cambodia: A set of directives called 'Prakas' were introduced in May 2018 by authorities in the run up to the general election. These directives enabled the government to regulate news websites and social media in the country.[107] In 2019 the Information Ministry announced renewed warnings to revoke the licenses of media outlets if 'found guilty of spreading disinformation that threatened 'national security'.[108]

China: The Criminal Law of the PRC states that anyone who makes up or knowingly spreads any false information about any risk, epidemic disease, disaster or emergency and spreads such information on the information network or any other media, which seriously disrupts public order, shall be sentenced to imprisonment of not more than three years, limited incarceration or surveillance. For serious circumstances, imprisonment may be seven years at most.[109] The Cybersecurity Law states that any individual or organization using the network shall not use the network to fabricate or disseminate false information to disrupt the economic and social order. Network operators shall be responsible for the management of information released by their users.[110] Besides, the Public Security Administration Punishment Law sets out that anyone who intentionally disturbs public order by spreading any rumour, giving false information about the situation of any risk, epidemic disease or emergency shall be detained for not more than 10 days and may concurrently be subject to a fine.[111]

Beyond legislative measures, China also deploys administrative measures to police disinformation. The Cyberspace Administration of China was re-organized in 2014 and was authorized by the State Council to be responsible for national Internet information content management as well as supervision, administration and law enforcement.[112] In its *Provisions on Ecological Governance of Network Information Content* published in 2019, network information content producers shall not make, copy or publish any illegal information which spreads rumours to disturb economic and social order.[113] Besides, the Cyberspace Administration of China set up the official website of 'China's Internet Joint Rumour Dispelling Platform'[114] for fact-checking nationwide. Each province also establishes their own official platforms mainly for checking rumours and fake news within the territorial scope of provinces. During the COVID-19 pandemic, the government cooperated with large tech companies (mainly including Tencent, Alibaba, ByteDance and Baidu) to provide the public with continuously updated fact-checking platforms concerning disinformation about public health, public policy, COVID-19 etc.[115]

---

106 Bangladesh Parliament, Act No 46 of the Year 2018, Dhaka, 23 Ashwin, 1426/08 October, 2018. Section 25 https://perma.cc/73HR-RNRE accessed 15 February 2021

107 Public Media Alliance, 'The rise of 'fake news' laws across South East Asia', 6 December 2019, https://www.publicmediaalliance.org/the-rise-of-fake-news-laws-across-south-east-asia/ accessed 15 February 2021

108 Wiseman, Jamie, 'Cambodia 'fake news' laws tighten noose on press freedom', International Press Institute, 1 October 2019, https://ipi.media/cambodia-fake-news-laws-tighten-noose-on-press-freedom/ accessed 15 February 2021

109 Standing Committee of the National People's Congress, Amendment (IX) to the Criminal Law of the People's Republic of China, 29 August 2015, available at http://search.chinalaw.gov.cn/law/searchTitleDetail?LawID=333125&Query=刑法 &IsExact=&PageIndex=4 accessed 15 February 2021

110 Standing Committee of the National People's Congress, the Cybersecurity Law of the People's Republic of China, 1 June 2017, available at http://search.chinalaw.gov.cn/law/searchTitleDetail?LawID=394858&Query=网络安全法&IsExact= accessed 15 February 2021

111 Standing Committee of the National People's Congress, the Public Security Administration Punishment Law, 1 January 2013, available at http://search.chinalaw.gov.cn/law/searchTitleDetail?LawID=333226&Query=治安管理处罚法&IsExact= accessed 15 February 2021

112 The State Council of the People's Republic of China, 26 August 2014, available at http://www.gov.cn/zhengce/content/2014-08/28/content_9056.htm accessed 15 February 2021

113 The Cyberspace Administration, Provisions on Ecological Governance of Network Information Content, 15 December 2019, available at http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm accessed 15 February 2021

114 Internet Joint Rumor Dispelling Platform, available at https://www.piyao.org.cn accessed 15 February 2021

115 For example, available at https://vp.fact.qq.com/home; https://mbd.baidu.com/newspage/data/mdpage?tag=8&id=5807 accessed 15 February 2021

India: Governments in India have used internet shutdowns to deal with fake news on several occasions, with one hundred reported shutdowns in 2018 alone.[116] In late February 2019 the Government of India released a Draft National E-Commerce Policy which notes that online platforms have a *'responsibility and liability . . . to ensure genuineness of any information posted on their websites.*'[117]

Indonesia: Indonesia's controversial Information and Electronic Transactions Act (UU ITE) has been accused of becoming the country's de facto 'anti-fake news' law.[118] The UU ITE permits the arrest of 'hoax news' distributors, however it has been admitted that some of these distributors shared content they believed was real.[119]A new website called Stophoax.id was also recently launched by the Ministry of Communication and Information Technology to bust hoaxes and provide an avenue for the public to report alleged sources of fake news.[120]

Japan: Japan's Broadcasting Act establishes a system to keep broadcasting programs from distorting facts and states that when a broadcaster edits broadcast programs, it must ensure the reporting does not distort information.[121] Additionally, under the Public Offices Election Act, anyone who conducts an election campaign online must make their online contact information available for users.[122] The purpose of this provision is to reduce the circulation of defamatory information and 'spoofing'. The Act on the Limitation of Liability of Internet Providers exempts internet providers from liability when they prevent distribution of information.[123] If there is a reasonable ground for a provider to believe that the rights of others were infringed without due cause by distribution of the information via a provider' service, the provider may block the information, and is not liable for the conduct of the person who sent the information.[124]

Pakistan: In 2020, Citizens Protection (Against Online Harm) Rules[125] came into force. Rule 4(1) sets an obligation for social media companies to block and remove unlawful online content within twenty-four hours following intimation by the national authority.[126] Rule 4(4) provides that social media companies should 'deploy proactive mechanisms' to prevent live streaming of content in violation of laws or instructions of the National Coordinator related to themes such as terrorism, hate speech, fake news, incitement to violence and national security.[127]

Singapore: The Protection from Online Falsehoods and Manipulation Act, or 'POFMA'[128] has been restricting freedom of expression by allowing the government to declare content 'false' and order its

116 'Government Responses To Disinformation On Social Media Platforms'. Loc.Gov, 2020, https://www.loc.gov/law/help/social-media-disinformation/india.php. Accessed 15 Feb 2021.
117 'Dipp.Gov.In'. Perma.Cc, 2019, https://perma.cc/C4H5-E86R. Accessed 11 Feb 2021. p.29
118 Tapsell, Ross, 'Indonesia's Policing of Hoax News Increasingly Politicised', ISEAS Yusof Ishak Institute, Perspective 2019(5), 20 September 2019, p. 2 https://www.iseas.edu.sg/images/pdf/ISEAS_Perspective_2019_75.pdf accessed 15 February 2021
119 Tapsell, Ross, 'Indonesia's Policing of Hoax News Increasingly Politicised', ISEAS Yusof Ishak Institute, Perspective 2019(5), 20 September 2019, p. 3 https://www.iseas.edu.sg/images/pdf/ISEAS_Perspective_2019_75.pdf accessed 15 February 2021
120 Sundari, Eva Kusuma, 'The fight against 'fake news' in Indonesia', Asia Times, 12 April 2019, https://asiatimes.com/2019/04/the-fight-against-fake-news-in-indonesia/ accessed 15 February 2021
121 Library of Congress, 'Iniatives to Counter Fake News: Japan', Index of Iniatives to Counter Fake News, April 2019, https://www.loc.gov/law/help/fake-news/japan.php, accessed 15/2/2021
122 Library of Congress, 'Iniatives to Counter Fake News: Japan', Index of Iniatives to Counter Fake News, April 2019, https://www.loc.gov/law/help/fake-news/japan.php, accessed 15/2/2021
123 Library of Congress, 'Iniatives to Counter Fake News: Japan', Index of Iniatives to Counter Fake News, April 2019, https://www.loc.gov/law/help/fake-news/japan.php, accessed 15/2/2021
124 Library of Congress, 'Iniatives to Counter Fake News: Japan', Index of Iniatives to Counter Fake News, April 2019, https://www.loc.gov/law/help/fake-news/japan.php, accessed 15/2/2021
125 Pakistan Ministry of Information Technology and Telecommunication, 21 January 2020, https://www.medianama.com/wp-content/uploads/CP_Against_Online_Harm_Rules_2020.pdf, accessed 15/2/2021
126 Pakistan Ministry of Information Technology and Telecommunication, 21 January 2020, Rule 4(1), https://www.medianama.com/wp-content/uploads/CP_Against_Online_Harm_Rules_2020.pdf, accessed 15/2/2021
127 Pakistan Ministry of Information Technology and Telecommunication, 21 January 2020, Rule 4(4), https://www.medianama.com/wp-content/uploads/CP_Against_Online_Harm_Rules_2020.pdf, accessed 15/2/2021
128 Republic of Singapore, 'Protection from Online Falsehoods and Manipulation Act 2019', No. 18 of 2019, date of commencement 2 October 2019, https://sso.agc.gov.sg/Act/POFMA2019#P12- accessed 15 February 2021

correction.[129] As of July 2020, POFMA had been invoked 55 times,[130] primarily on content that was critical of the government and its policies. [131] Orders of correction have been sent to the independent online media, opposition politicians,[132] NGOs,[133] and activists.[134] Non-compliance with POFMA may lead to fines and jail time.[135]

Thailand: The Digital Economy and Society (DES) Ministry opened an anti-fake-news centre to verify the truth and give feedback to citizens.[136] It will have two committees to support it, the first made up of academics and experts and the second of representatives from various organisations, both governmental and non-governmental.[137]

United Arab Emirates: The Federal Telecommunication Regulatory Authority instructs internet service providers to block any online content promoting violence, pornography, and political speech. In 2017, the Authority blocked several Qatari media websites, including Al-Jazeera Live, Peninsula Qatar, the Arabic Huffington Post, and the Muslim Brotherhood's Official website. Furthermore, Article 38 of the Federal Law No. 5 of 2012 stipulates that whoever publishes online any incorrect, inaccurate, or misleading information that damages the interests of the State or tarnishes its reputation, prestige, or stature must be punished with a term of imprisonment.[138] Under article 39, any person who fails to remove or block access to illicit content after receiving a notice from the federal authorities faces a term of imprisonment, a fine, or both.

Vietnam: the cybersecurity law of 2019, placing stringent controls on tech firms including setting up offices in the country, storing data locally and complying with Hanoi's demands to delete anti-state' content[139] on social media.[140]

IV. Africa

129 Human Rights Watch, 'Singapore: Events of 2020, https://www.hrw.org/world-report/2021/country-chapters/singapore, accessed 15/2/2021.
130 Meyer, Paul, 'Singapore's First Election Under the Fake News Law' The Diplomat, 7 July 2020, https://thediplomat.com/2020/07/singapores-first-election-under-the-fake-news-law/ accessed 15 February 2021
131 Human Rights Watch, World Report 2021, Singapore https://www.hrw.org/world-report/2021/country-chapters/singapore accessed 15 February 2021
132 John Tan, opposition leader, was found guilty under POFMA
Abu Baker, Jalelah, 'Jolovan Wham, SDP's John Tan fined S$5,000 for contempt of court', CNA, 29 April, 2019, https://www.channelnewsasia.com/news/singapore/jolovan-wham-sdp-john-tan-fined-contempt-of-court-11487364, accessed 15/2/2021
133 Lawyers for Liberty's website was blocked after the NGO did not comply with a Correction Order; Ministry of Communications and Information, 'Minister for Communications and Information Directs IMDA to Issue Access Blocking Orders, 23 January 2020, https://www.pofmaoffice.gov.sg/documents/media-releases/2020/January/mci-imda-abo-23-jan.pdf, accessed 15/2/2021
134 Activist Jolovan Wham was found guilty of contempt of court over a Facebook post, Cna, 09 Oct 2018 02:08PM https://www.channelnewsasia.com/news/singapore/activist-jolovan-wham-found-guilty-of-scandalising-judiciary-10806894 accessed 15 February 2021
135 Human Rights Watch, World Report 2021, Singapore, available at https://www.hrw.org/world-report/2021/country-chapters/singapore accessed 15 February 2021
136 Suchit Leesa-nguansuk, Centre goes live to fight fake news, Bangkok Post, 2 NOV 2019 available at https://www.bangkokpost.com/business/1785199/centre-goes-live-to-fight-fake-news accessed 15 February 2021
137 Post Reporters, Govt's anti-fake news centre gets help, Bangkok Post, 18 Feb 2020, Available at https://www.bangkokpost.com/thailand/general/1859719/govts-anti-fake-news-centre-gets-hel
138 Federal Law No. 5 of 2012, Al-Jaridah Al-Rasmiyah, vol. 540 (13 Aug. 2012),
139 Sonia Sarkar, Vietnam artists seek 'liberation' from cybersecurity law, DW, 18.01.2019, https://www.dw.com/en/vietnam-artists-seek-liberation-from-cybersecurity-law/a-47119106 accessed 15 February 2021
140 Emmy Sasipornkarn, Southeast Asia 'fake news' laws open the door to digital authoritarianism, DW, 16.10.2019, https://www.dw.com/en/southeast-asia-fake-news-laws-open-the-door-to-digital-authoritarianism/a-50852994 Accessed 15 February2021

Cameroon: In April 2020, the National Agency for Information and Communication Technologies sent SMS messages to cell phone users in the country which warned about penalties for spreading false news.[141] These SMS messages contained a reminder of the penalties for violating Article 78(1) of Law N°2010/012 of 21 December 2010 on Cybersecurity and Cybercrime in Cameroon.[142]

Ethiopia: Ethiopia has a specific law to tackle disinformation, the Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185/2020.[143] Article 5 broadly criminalises the dissemination of disinformation, which Article 2 defines as, 'speech that is false, is disseminated by a person who knew or should reasonably have known the falsity of the information and is highly likely to cause a public disturbance, riot, violence or conflict'. Furthermore, Article 8(1) requires social media service providers to 'endeavour to suppress and prevent the dissemination of disinformation' on their platforms, while Article 8(2) requires them to act within twenty-four hours to remove disinformation on their platforms upon receiving notification of its existence. While aimed at curbing historical tensions in relation to hate speech and violence in Ethiopia, this law presents several issues from a human rights perspective.[144] Its scope is broadly defined, meaning that the authorities could potentially interpret this law as giving them the power to restrict a wide range of speech.[145] The applicable penalties under Article 7 are potentially disproportionate in their severity,[146] and may have a chilling effect on freedom of expression.[147]

Kenya: Under the Computer Misuse and Cybercrimes Act, section 22 prohibits individuals from intentionally publishing false, misleading or fictitious data or misinforming with intent that the data be considered or acted upon as authentic, and section 23 prohibits individuals from knowingly publishing false information in print, broadcast, data or over a computer system, which 'is calculated or results in panic, chaos, or violence among citizens of the Republic, or which is likely to discredit the reputation of a person'.[148] These provisions are broadly construed, and it is unclear how to determine the scope of what is considered 'false'.[149]

Nigeria: Nigeria currently has a proposal for legislation that aims to counter disinformation: the Protection from Internet Falsehoods and Manipulation and Other Related Matters Bill 2019.[150] Additionally, Nigeria has two laws that include restrictions on disinformation, the Cybercrimes (Prohibition, Prevention, etc) Act 2015, and the Criminal Code, 1990.[151] The proposed law and the current legislation raise significant concerns in relation to the freedom of expression.[152] They are broadly defined in their scope, which means that the authorities could use them to restrict a wide range of speech. For

141'Disinformation Tracker'. Disinformationtracker.Org, https://www.disinformationtracker.org/. Accessed 15 Feb 2021.
142 Ibid.
143 Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185/2020, https://chilot.me/wp-content/uploads/2020/04/HATE-SPEECH-AND-DISINFORMATION-PREVENTION-AND-SUPPRESSION-PROCLAMATION.pdf, accessed 14 February 2021
144 'Disinformation Tracker'. Disinformationtracker.Org, https://www.disinformationtracker.org/. Accessed 15 Feb 2021.
145 Article 19, 'Ethiopia: Hate speech and disinformation law must not be used to suppress the criticism of the Government' (19 January 2021), https://www.article19.org/resources/ethiopia-hate-speech-and-disinformation-law-must-not-be-used-to-supress-the-criticism-of-the-government/, accessed 13 January 2021
146 Under Article 7 of the Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185/2020, any person who disseminates disinformation may be punished with simple imprisonment not exceeding one year or a fine not exceeding 50,000 birr; if the offense has been committed through a social media account having more than 5,000 followers or through a broadcast service or print media, there is a penalty of simple imprisonment not exceeding three years or a fine not exceeding 100,000 birr; and if violence or public disturbance occurs due to the dissemination of disinformation, the punishment shall be rigorous imprisonment from two years up to five years.
147 'Disinformation Tracker'. Disinformationtracker.Org, https://www.disinformationtracker.org/. Accessed 15 Feb 2021.
148 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
149 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
150 Protection from Internet Falsehoods and Manipulation and Other Related Matters Bill 2019, <https://www.nassnig.org/documents/billdownload/10965.pdf> accessed 12 February 2021.
151 Cybercrimes (Prohibition, Prevention, etc) Act 2015 <https://www.cert.gov.ng/ngcert/resources/CyberCrime__Prohibition_Prevention_etc__Act__2015.pdf> accessed 12 February 2021; the Criminal Code, 1990 <https://www.wipo.int/edocs/lexdocs/laws/en/ng/ng025en.pdf> accessed 13 February 2021.
152 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.

example, section 24(b) of the Cybercrimes (Prohibition, Prevention, etc) Act 2015, creates a criminal offence of knowingly or intentionally publishing a message online when the individual knows the message 'to be false, for the purpose of causing annoyance, inconvenience, danger, obstruction, insult, injury, criminal intimidation, enmity, hatred, ill will or needless anxiety to another or causes such a message to be sent'. It is unclear how to determine whether a message is 'false', or what is included under the scope of 'causing annoyance, inconvenience, danger, obstruction, insult, injury, criminal intimidation, enmity, hatred, ill will or needless anxiety to another'.[153] In 2020, the ECOWAS (Economic Community of West African States) Court decided that elements of the Cybercrimes Act violated the right to freedom of expression under regional and international human rights law and ordered the Nigerian government to either appeal or amend it.[154] However, the government has not acted on this so far.[155] Furthermore, these laws pursue aims which are not considered legitimate under international human rights standards—for example, restricting speech which might be prejudicial to 'public tranquillity or public finances'.[156] These laws also carry penalties which risk being disproportionate and resulting in a chilling effect on the freedom of expression in Nigeria.[157]

Senegal: Article 255 of the Penal Code criminalises the publication, dissemination, disclosure or reproduction of false news ('nouvelles fausses') when it causes or is likely to cause disobedience of the country's laws, damage to the morale of the population or discredits public institutions.[158] However, it is not clear how to determine whether news is 'false', and what threshold is required for damaging public morale or discrediting public institutions. [159] In fact, restrictions in pursuance of avoiding damage to public morale or bringing public institutions into disrepute are outside the scope of what is considered 'public order' and therefore constitute illegitimate objectives.[160]

South Africa: South Africa does not currently have specific legislation to counter disinformation. However, there are laws and proposed laws that include restrictions on certain forms of disinformation: the Regulations issued in terms of Section 27(2) of the Disaster Management Act, 2002, and the Cybercrimes and Cybersecurity Bill, 2017.[161] Both raise substantial concerns from a human rights perspective. Firstly, their scope is ill-defined, meaning that the Government could penalise a broad range of expression, and they may pursue aims which would not be considered 'legitimate' according to international human rights standards—for example, restricting speech which might cause psychological or economic harm (Cybercrimes and Cybersecurity Bill, 2017).[162] For example, section 11(5) of the Regulations issued in terms of Section 27(2) of the Disaster Management Act criminalises the publication of any statement made 'with the intention to deceive any other person' about COVID-19, the infection status of any person, or any measure taken by the government to address the pandemic. Notably, this could potentially be used by the authorities to restrict speech which is critical of government measures. In terms of the Cybercrimes and Cybersecurity Bill, 2017, section 17(2)(d) criminalises the distribution of any data message that is harmful, including messages that are 'inherently false in nature' and 'aimed at causing mental, psychological,

153 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
154 Innocent Odoh, 'ECOWAS Court orders Nigeria to amend its law on cybercrime' (Business Day, 11 July 2020) <https://businessday.ng/security/article/ecowas-court-orders-nigeria-to-amend-its-law-on-cybercrime/#:~:text=The%20ECOWAS%20Court%20of%20Justice,on%20Civil%20and%20Political%20Rights.> accessed 14 February 2021.
155 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
156 Protection from Internet Falsehoods and Manipulation and Other Related Matters Bill 2019, section 3(1)(b); Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
157 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
158 Penal Code <http://www.droit-afrique.com/upload/doc/senegal/Senegal-Code-1965-penal.pdf> accessed 14 February 2021.
159 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
160 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
161 Regulations issued in terms of Section 27(2) of the Disaster Management Act, 2002 <https://www.gov.za/sites/default/files/gcis_document/202003/43107gon318.pdf> accessed 13 February 2021. Cybercrimes and Cybersecurity Bill, 2017 <https://www.justice.gov.za/legislation/bills/CyberCrimesBill2017.pdf> accessed 13 February 2021; Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
162 Cybercrimes and Cybersecurity Bill, 2017, section 17(2)(d); Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.

physical or economic harm to a specific person or group of persons'. Additional guidance is needed on what the scope of psychological harm, and what type of economic harms, whether individual, company, or state level, are within scope.[163]

Zimbabwe: Zimbabwe currently has no specific legislation on disinformation. However, there are several laws that include potential restrictions on disinformation: the Cybercrime and Cybersecurity Bill, 2017, the Criminal Law (Codification and Reform) Act, and Statutory Instrument (SI) 83 of 2020, the Public Health (COVID-19 Prevention, Containment and Treatment) (National Lockdown) Order, 2020.[164] All three instruments raise significant concerns from a human rights perspective. They are broadly defined in scope, which means that authorities could invoke them to restrict a wide range of expression.[165] For example, the Cybercrime and Cyber Security Bill, 2017 includes a provision that criminalises the transmission of a 'false data message intending to cause harm'.[166] Under section 17, 'any person who unlawfully and intentionally by means of a computer or information system makes available, broadcasts or distributes data to any other person concerning an identified or identifiable person knowing it to be false with intent to cause psychological or economic harm shall be guilty of an offence'. It is not clear how it would be determined that a message was 'false' or what 'psychological or economic harm' entails.[167] Furthermore, these laws pursue aims which are not considered legitimate according to international human rights standards—for example, restricting speech which might adversely affect the economic interests of the country (Criminal Law (Codification and Reform) Act).[168]

2(b). *What has been the impact of such measures on i) disinformation; ii) freedom of opinion and expression; and iii) other human rights?*

Several of the initiatives to counter disinformation have been criticized for curtailing freedom of expression; freedom of media/press; liberty and security by civil society and national and international organizations as well as NGOs. For example, the UN High Commissioner for Human Rights, Michelle Bachelet, expressed concern in April 2020 over the restrictive measures imposed by several States against the independent media, as well as the arrest and intimidation of journalists in the context of the covid-19 pandemic.[169]

I. Europe

In Europe, the Council of Europe's Commissioner for Human Rights alerted that legislation countering disinformation adopted in several member states was curtailing the work of journalists and restricting the public's access to information. The Commissioner expressed concern over several of the decrees and laws passed in Hungary, Russia, Azerbaijan, Romania, Bosnia and Herzegovina and Armenia as well as reprisal actions against journalists.[170] Legislation in Hungary has received the harshest criticism given the increased

163 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
164 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
165 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
166 Cybercrime and Cyber Security Bill, 2017, section 17.
167 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
168 Criminal Law (Codification and Reform) Act, section 31(a)(ii); Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.
169 UN Office of the High Commissioner for Human Rights,
Bachelet alarmed by media clampdowns, says public has right to know about COVID-19, 24 April 2020, available at https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25823&LangID=E
170 CoE Commissioner for Human Rights, Press freedom must not be undermined by measures to counter disinformation about COVID-19, 3 April 2020, available at https://www.coe.int/en/web/commissioner/-/press-freedom-must-not-be-undermined-by-measures-to-counter-disinformation-about-covid-19 accessed 15 February 2021

risk of criminal prosecution of journalists and the chilling effect on media and freedom of expression.[171] In Serbia, a governmental decree enforcing the centralization of COVID-19 information and imposing sanctions for local institutions if they released information without authorization from the capital was struck down following intense public criticism.[172] A website on COVID-19 fake news launched by the French government (Desinfox) was also struck down after an emergency appeal to the Conseil d'État by the State's journalists' union alleging that the page was a 'clear interference by public authorities in the freedom of press.'[173]

II. The Americas

In the Americas, the most worrying development has been with regards to the 'Brazilian Internet Freedom, Responsibility and Transparency Act'. Several organizations such as Human Rights Watch,[174] Article 19[175] and the Global Network Initiative[176] have alerted that this bill poses a serious threat to freedom of expression and privacy in Brazil. The main concerns raised are the onerous 'traceability' obligations which impose significant data retention requirements on private messaging services and the provisions requiring official documents to users. A joint declaration signed by several Brazilian NGOs, including Amnesty International, has also stated that the bull fails to comply with the goal of combatting disinformation, given that it stimulates a very concentrated digital environment by imposing burdensome obligations on internet service providers, encouraging censorship and creating a "chilling effect" on online freedom of expression[177]. Bolivia's COVID-19 decree has also sparked criticism for containing overly vague provisions criminalizing disinformation which open the possibility of abuse against journalists.[178]

III. Asia-Pacific

In India the authorities have charged journalists and one doctor for their public criticism of the Government's response to the pandemic. To a certain extent this has led to the self-censorship by various Indian news outlets which retracted its articles on COVID-19 without any explanation.[179] Many former Supreme Court judges, writers, chiefs of naval staffs have endorsed a statement criticising the Government's statements.[180] Furthermore, the Government has tried to restrict the access to information by demanding

171 International Press Institute, Hungary seeks power to jail journalists for 'false' COVID-19 coverage, March 23 2020, available at https://ipi.media/hungary-seeks-power-to-jail-journalists-for-false-covid-19-coverage/ accessed 15 February 2021

172 Radu, Roxana. (2020). Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation. Social Media Society, 6(3), p. 2.

173 The Guardian, The French government takes down coronavirus 'fake news' web page, 6 May 2020, available at https://www.theguardian.com/world/2020/may/06/french-government-takes-down-coronavirus-fake-news-web-page accessed 15 February 2021

174 Human Rights Watch, Brazil: Reject 'Fake News' Bill, June 24 2020, available at https://www.hrw.org/news/2020/06/24/brazil-reject-fake-news-bill accessed 15 February 2021

175 Article 19, Brazil: Fake new law threatens freedom of expression, 3 July 2020, available at https://www.article19.org/resources/fake-news-law-threatens-freedom-of-expression-in-brazil/ accessed 15 February 2021

176 Global Network Initiative, GNI Expresses Concern About Proposed 'Fake News' Law in Brazil, June 29 2020, available at https://globalnetworkinitiative.org/gni-concerns-brazil-fake-news-law/ accessed 15 February 2021

177 Freedom House, Joint Statement, Brazil: Disinformation Bill Threatens Freedom of Expression and Privacy Online, June 29 2020, available at https://freedomhouse.org/article/brazil-disinformation-bill-threatens-freedom-expression-and-privacy-online accessed 15 February 2021

178 Committee to Protect Journalists, Bolivia enacts decree criminalizing 'disinformation' on COVID-19 outbreak, April 9 2020, available at https://cpj.org/2020/04/bolivia-enacts-decree-criminalizing-disinformation/ accessed 15 February 2021 see also Human Rights Watch, Bolivia: COVID-19 Decree Threatens Free Expression, April 7 2020, available at https://www.hrw.org/news/2020/04/07/bolivia-covid-19-decree-threatens-free-expression, accessed 15 February 2021

179 Bansari Kamdar, COVID-19 and Shrinking Press Freedom in India, The Diplomat, 29 May 2020, available at https://thediplomat.com/2020/05/covid-19-and-shrinking-press-freedom-in-india/ accessed 15 February 2021

180 ibid

journalists to publish only 'official information'.[181] At the same time, the Supreme was requested to 'issue a direction that no electronic/print media/web portal or social media shall print/publish or telecast anything without first ascertaining the true factual position from the separate mechanism provided by the Government'. [182] However, the Supreme Court denied the petition, but did direct the media to 'refer to an publish the official version of the developments'.[183] Such trends can lead to the curtailing of the right to liberty and security as well. For example, in Bangladesh several people, including journalists and human rights activists, have been charged or arrested under the Digital Security Act for 'spreading disinformation about Covid-19 or criticising the Government's response.'[184] For instance, on 5 May 2020 human rights activist Didar Bhuyian was arrested for criticising the Government's response to the pandemic.[185]

In China, according to the UN Human Rights Office, Chinese Authorities have detained, and in some cases, charged medical professionals, academics, and ordinary citizens for 'publishing their views or information related to COVID-19, or who have been critical of the government.'[186] For example, Chinese officials punished eight whistle-blowers for 'spreading rumours' about the new virus and 'disturbing the social order' in the early days of the outbreak.[187] They were arrested by the police and punished after sharing their views with friends through the social media platform, WeChat. This practice was soon later criticized by the Chinese Supreme People's Court, which said the eight citizens should have not been punished because what they spread was not entire false information and would help people to carry out sanitization measures at the early stage.[188] Similar observations have been made in Nepal, Indonesia, the Philippines, and Malaysia.[189]

IV. Africa

In Africa similar observations have been made where the criminalisation of disinformation has led to the curtailing of freedom and expression and liberty and security. For instance, in Ethiopia Yayesew Shimelis, a journalist, was arrested in April 2020 and charged with violating the Hate Speech and Disinformation Prevention and Suppression Proclamation for sharing a Facebook post that suggested the government had prepared 200,000 burial places in response to COVID-19.[190] Shimelis was detained for three weeks before any charges were brought against him, violating his right to liberty.[191] In Botswana, three individuals, including the opposition spokesperson Justice Motlhabane, were arrested and charged for the publication

181 Radu, R. (2020). Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation. Social Media + Society, 6(3), p.205630512094819. available at https://journals.sagepub.com/doi/pdf/10.1177/2056305120948190 accessed 15 February 20212

182 ibid ; see also: PTI, Centre seeks SC direction, says no media should publish Covid-19 info before checking facts with govt, The Times of India, Mar 31, 2021 available at HYPERLINK 'https://timesofindia.indiatimes.com/india/centre-in-sc-no-media-should-publish-covid-19-info-without-ascertaining-facts-with-govt/articleshow/74918103.cms'https://timesofindia.indiatimes.com/india/centre-in-sc-no-media-should-publish-covid-19-info-without-ascertaining-facts-with-govt/articleshow/74918103.cms accessed 15 February 2021

183 Bansari Kamdar, COVID-19 and Shrinking Press Freedom in India, The Diplomat, 29 May 2020, available at https://thediplomat.com/2020/05/covid-19-and-shrinking-press-freedom-in-india/ accessed 15 February 2021

184 Asia: Bachelet alarmed by clampdown on freedom of expression during COVID-19, 3 June 2020 https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25920

185 https://www.frontlinedefenders.org/en/statement-report/two-years-coming-force-bangladeshs-digital-security-act-continues-target-human

186 https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25920

187 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30382-2/fulltext

188 https://asia.nikkei.com/Spotlight/Caixin/Rebuked-coronavirus-whistleblower-vindicated-by-top-Chinese-court

189 https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25920

190'News Update: Prosecutors Charge Journalist Yayesew with Newly Enacted Hate Speech Law' (21 April 2020) <https://addisstandard.com/news-update-prosecutors-charge-journalist-yayesew-with-newly-enacted-hate-speech-law/> accessed 12 February 2021.

191 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021.

of fake news related to COVID-19.[192] Two of the articles suggested that the President had declared a lengthy state of emergency so that he could deal with his political rivals and business competitors, while another article questioned why individuals infected with COVID-19 in hospital were not developing further complications or recovering.[193] A police spokesperson claimed that these three men had published an 'offensive statement against the government' as well as 'degrading and maligning the leadership of the country'. However, these individuals and their lawyers argue that the arrests were politically motivated, and that the government is criminalising legitimate expression.[194]

Hence, it is important to stress that when actions are used to control the flow of information to counter disinformation it can be dangerous and often counterproductive. Especially, when information is related to general interests like public health, it will prevent both citizens and authorities from responding quickly to avoid further serious consequences.

2(c). What Measures Have Been Taken To Address Any Negative Impact On Human Rights?

Several measures have been taken to address negative impact on human rights at the national; regional; and international level. These are primarily made up of soft law mechanisms and guidance.

I. International level

At the UN level, the Special Rapporteur had published reports on State obligation to safeguard human rights in the line of taking measures against disinformation. The SR has acknowledged that measures on disinformation can lead to curtailing human rights. For example, regarding Artificial Intelligence (AI), the SR noted that AI-driven personalization has reinforced biases and has incentivized the promotion and recommendation of inflammatory content or disinformation to sustain online engagement.[195] Hence, States should ensure that the development of AI is in line with Article 17, 19 and 26 ICCPR, and devise national AI policies to explore and develop strategies for the maximum benefit for all their citizens.[196] However, the SR stressed that the AI could interfere with the right to effective remedies; therefore, States must make effective remedies available to individuals. Individuals should not only be made aware that they have been subject to an algorithmic decision, but they should also be equipped with information about the reason behind the decision.[197]

Regarding online hate speech and disinformation, the SR has stressed that States have an international obligation to regulate online hate speech. However, any limitation to online expression must meet the standards under Article 19(3) and Article 20 of International Covenant on Civil and Political Rights (ICCPR)

192 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021; Pako Lebanna, 'Court remands trio in custody' (Daily News Botswana, 14 April 2020) <http://www.dailynews.gov.bw/news-details.php?nid=55626> accessed 14 February 2021.

193 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021; Joel Konopo, 'Censorship, the unexpected side-effect of Covid-19' <https://mg.co.za/africa/2020-05-11-censorship-the-unexpected-side-effect-of-covid-19/> accessed 14 February 2021.

194 Disinformation Tracker <https://www.disinformationtracker.org/> accessed 14 February 2021; Joel Konopo, 'Censorship, the unexpected side-effect of Covid-19' <https://mg.co.za/africa/2020-05-11-censorship-the-unexpected-side-effect-of-covid-19/> accessed 14 February 2021.

195 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/73/348, 29 August 2018, para. 12.

196 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/73/348, 29 August 2018, para. 42.

197 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/73/348, 29 August 2018, para. 60.

and Article 4 of International Convention on Elimination of All forms of Racial Discrimination (ICERD).[198] This means that any limitation must be prescribed by law; serve a legitimate aim; and is necessary and proportional to the means to achieve the aim in a democratic society.[199] Hence, States need to tie their definition of hate speech to the standards of international human rights framework; otherwise, it may lead to risk of abuse, restriction of legitimate content and failure to address the issue at hand.[200]
Additionally, there should be adequate safeguards against arbitrariness - including the right to appeal.[201] The aim for restricting or limiting the right must be justified under Article 19(3) ICCPR.[202] Lastly, any legislative effort to remove online hate speech and impose liability on platforms must meet the necessity and proportionality requirements: it is vital that limitations are applied o for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicted.[203]

In fighting the infodemic crisis during COVID-19, the SR highlighted the importance of the 2017 joint Declaration on freedom of expression and 'fake news', disinformation and propaganda. The declaration included: (i) State actors should not make, sponsor, encourage or further disseminate statements which they know or reasonably should know to be false (disinformation), or which demonstrate a reckless disregard for verifiable information (propaganda); (ii) State actors should, in accordance with their domestic and international legal obligations and their public duties, take care to ensure that they disseminate reliable and trustworthy information, including about matters of public interest, such as the economy, public health, security and the environment.[204] The Declaration made clear that general prohibitions on the dissemination of information based on 'vague and ambiguous ideas, including 'false news' or 'non-objective information' are incompatible with human rights law and should be abolished'.[205]Additionally, 'vague prohibitions of disinformation effectively empower government officials with the ability to determine the truthfulness or falsity of content in the public and political domain, in conflict with the requirements of necessity and proportionality under article 19 (3).'[206]

II. Regional level

a. African soft law and guidance

At the African human rights regional level, in 2019 the Declaration of Principles on Freedom of Expression and Access to Information was adopted.[207] Under Principle 9 States may only limit the exercise of the rights to freedom of expression and access to information, if the limitation is: prescribed by law; serve a legitimate aim; and is necessary and proportionate means to achieve the stated aim in a democratic society.[208] Furthermore, any law limiting this right must be clear, precise, accessible, and foreseeable; overseen by an independent body in a manner that is not arbitrary or discriminatory; and effectively safeguard against

198 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 28.
199 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 6.
200 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 31.
201 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 7, 33.
202 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 39.
203 United Nations, 'Promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para. 6.
204 United Nations, 'Disease Pandemics and the freedom of opinion and expression', A/HRC/44/49, 23 April 2020, para. 44.
205 United Nations, 'Disease Pandemics and the freedom of opinion and expression', A/HRC/44/49, 23 April 2020, para. 49.
206 United Nations, 'Disease Pandemics and the freedom of opinion and expression', A/HRC/44/49, 23 April 2020, para. 49.
207 Declaration of Principles on Freedom of Expression and Access to Information in Africa 2019 <https://www.achpr.org/legalinstruments/detail?id=69> accessed 14 February 2021; International Justice Resource Centre, 'New ACHPR Declaration on Freedom of Expression and Access to Information' (22 April 2020) <https://ijrcenter.org/2020/04/22/new-achpr-declaration-on-freedom-of-expression-access-to-information/> accessed 13 February 2021.
208 Declaration of Principles on Freedom of Expression and Access to Information in Africa 2019, Principle 9(1).

abuse including through the provision of a right of appeal to independent and impartial courts.[209] Finally, to be necessary and proportionate, the limitation shall originate from a pressing and substantial need that is relevant and sufficient; have a direct and immediate connection to the expression and disclosure of information and be the least restrictive means of achieving the stated aim; be such that the benefit of protecting the stated interest outweighs the harm to the expression and disclosure of information, including with respect to the sanctions authorized.[210] These principles should constrain African governments when they legislate to restrict disinformation. For example, Principal 22(2) provides that, 'States shall repeal laws that criminalize sedition, insult, and publication of false news', which means that laws criminalising the dissemination of disinformation are contrary to regional human rights norms. Furthermore, Principle 38(2) on non-interference provides that, 'States shall not engage in or condone any disruption of access to the internet and other digital technologies for segments of the public or an entire population', which is particularly relevant on the continent considering the rising number of internet shutdowns.[211] Lastly, Principle 38(4) provides that States shall not require the removal of online content by internet intermediaries unless such requests are, *inter alia*, clear and unambiguous, imposed by an independent and impartial judicial authority, justifiable and compatible with international human rights law and standards, and implemented through a transparent process that allows a right of appeal.[212]

Moreover, the African Declaration of Internet Rights and Freedoms is a pan-African initiative that seeks to promote human rights standards and principles of openness in internet policy formulation and implementation on the continent.[213] It does this through 13 principles which it views as necessary for upholding human rights online.[214] Under the Principle of Freedom of Expression, the Declaration states that, 'Content blocking, filtering, removal and other technical or legal limits on access to content constitute serious restrictions on freedom of expression and can only be justified if they strictly comply with international human rights law as reiterated in Article 3 of this Declaration'. Furthermore, the Declaration requires that, 'no-one should be held liable for content on the Internet of which they are not the author. To the extent that intermediaries operate within self-regulatory systems, and/or make judgement calls on content and privacy issues, all such decisions should be made considering the need to protect expression that is legitimate under the principles provided for under international human rights standards, including the Manila Principles on Intermediary Liability. Processes developed by intermediaries should be transparent and include provisions for appeals.' These principles provide key guidance to States on the limits of regulation, and the need to consider human rights as an integral part of any regulatory framework.

Finally, the Special Rapporteur on Freedom of Expression and Access to Information in Africa released a statement in April 2020 which provided that, 'internet and social media shutdowns violate the right to freedom of expression and access to information, contrary to Article 9 of the African Charter on Human and Peoples' Rights',[215] in line with Principle 38(2) in the Declaration of Principles on Freedom of Expression and Access to Information 2019. The Special Rapporteur called on African States to, 'take all measures to guarantee respect and protect the right to freedom of expression and access to information through ensuring access to internet and social media services especially during the COVID-19 pandemic.

209 Declaration of Principles on Freedom of Expression and Access to Information in Africa 2019, Principle 9(2).
210 Declaration of Principles on Freedom of Expression and Access to Information in Africa 2019, Principle 9(4).
211 Christopher Giles and Peter Mwai, 'Africa internet: Where and how are governments blocking it?' <https://www.bbc.com/news/world-africa-47734843> accessed 14 February 2021.
212 Christopher Giles and Peter Mwai, 'Africa internet: Where and how are governments blocking it?' <https://www.bbc.com/news/world-africa-47734843> accessed 14 February 2021.
213 African Declaration of Internet Rights and Freedoms <https://africaninternetrights.org/sites/default/files/African-Declaration-English-FINAL.pdf> accessed 13 February 2021.
214 APC, 'African Declaration on Internet Rights and Freedoms Coalition: Promotion of freedom of expression a priority for Southern Africa' (30 September 2019) <https://www.apc.org/en/news/african-declaration-internet-rights-and-freedoms-coalition-promotion-freedom-expression> accessed 13 February 2021.
215 'Press Release by the Special Rapporteur on Freedom of Expression and Access to Information in Africa on the Importance of Access to the Internet in Responding to the COVID-19 Pandemic' (8 April 2020) <https://www.achpr.org/pressrelease/detail?id=487> accessed 14 February 2021.

States should not disregard rule-of-law dictates by exploiting the pandemic to establish overreaching interventions.'[216]

b. Inter-American soft law and guidance

The 2017 joint Declaration was also signed by the Organization of American States.[217] According to the Declaration another focus of States' actions should be towards education to promote media and digital literacy, including by covering these topics as part of the regular school curriculum and by engaging with civil society and other stakeholders to raise awareness about these issues.[218]

The OAS also calls for the adoption of clear and transparent guidelines is prerequisite for content moderation by intermediaries. These actions include providing easy access to policies and rules, notification of any take down procedure and the possibility to contest the decision.

c. European soft law and guidance

At the EU level, the European Commission co-funds (together with the European Parliament) 'independent projects in the field of media freedom and pluralism.' These projects, among other actions, monitor risks to media pluralism across Europe, map violations to media freedom, fund cross-border investigative journalism and support journalists under threat.'[219]

3(a). What Policies, Procedures Or Other Measure Have Platforms Introduced To Address The Problem Of Disinformation?

Much of the public discussion concerning false information concerns the steps private companies should take to remove such information from platforms or punish those who spread false information. Platforms have been opposing reforms that would make them responsible for disinformation on their platforms.[220] However, as discussed in the earlier section, governments have made platforms liable for the spread of disinformation - if they do not act against it. Some platforms recognize their role in the spread of disinformation, and they have taken steps to help limit the phenomenon. Platform policies to address disinformation may be divided into three categories: **Interactive**, **Behavioural,** and **Restrictive** policies.

Interactive policies consist in outreach programs created by the platforms. These programs increase online literacy by actively engaging the users and by connecting them directly to experts. These policies aim to educated users in recognizing disinformation through preventative education.

Behavioural policies are aimed at changing user behaviour to encourage more interaction with content. This commonly includes placing a 'label'/'warning'/'banner' on posts having certain types of misleading, false, or harmful content. These labels may also direct users to trustworthy sources. Their aim is to educate those that meet disinformation and encourage more public discourse. That can also be said to boost user

216 'Press Release by the Special Rapporteur on Freedom of Expression and Access to Information in Africa on the Importance of Access to the Internet in Responding to the COVID-19 Pandemic' (8 April 2020)
<https://www.achpr.org/pressrelease/detail?id=487> accessed 14 February 2021.
217 Organization of American States, 'Joint Declaration on Freedom of Expression and 'Fake News', disinformation and propaganda', 2017, https://www.oas.org/en/iachr/expression/showarticle.asp?artID=1056&lID=1, accessed 15/2/2021
218 Organization of American States, 'Joint Declaration on Freedom of Expression and 'Fake News', disinformation and propaganda', 2017, https://www.oas.org/en/iachr/expression/showarticle.asp?artID=1056&lID=1, accessed 15/2/2021
219 Questions and Answers – Code of Practice against Disinformation: Commission Calls on Signatories to Intensify their Efforts, European Commission (Jan. 29, 2019)
https://www.loc.gov/law/help/social-media-disinformation/eu.php
220 Carp, Paul. 'Fake Online Covid Claims Should Be Exposed By Tech Companies, Health Experts Say'. The Guardian, 2021, https://www.theguardian.com/australia-news/2021/jan/25/tech-companies-should-be-forced-to-reveal-viral-covid-19-material-health-experts-say. Accessed 15 Feb 2021.

competences about understanding the quality of information, rather than nudge them toward popular, poor-quality news.

Restrictive polices consist in the 'de-amplification' of misleading, false, or harmful content. Their aim is to limit the visibility and the creation of misleading content. Such policies can include de-amplifying, deleting, or banning certain content, or deactivation of a user's account.

This section analyses what **interactive**, **behavioural,** and **restrictive** measures companies have used across a variety of topic areas. Topic discussed are **election/civic integrity**; **COVID-19**; **manipulated media**; **impersonation**; and **fake engagement** policies.

I. Election/civic integrity policies

The platforms surveyed have all created policies specific to civic processes, such as elections and political campaigns. Many of the policies focus on either directing users to more authoritative information or warning users of potentially false information, de-amplifying, or even removing content which has false or misleading information about election processes and outcomes. Rarer are interactive measures that work to engage users or educate on false news surrounding elections and political processes.

Election			
	Interactive policies	Behaviour policies	Restrictive policies
Facebook	☑	☑	☑
Twitter	☐	☑	☑
Reddit	☐	☑	☑
Youtube	☐	☐	☑
Google	☑	☑	☐
TikTok	☑	☐	☑
Amazon	☐	☐	☐
Snapchat	☐	☑	☐

a. Interactive policies

Facebook has started a campaign in Africa on certain radio stations and on Facebook to boost digital literacy by providing educational tips on how to spot fake news.[221] Although this is a part of Facebook's efforts to support elections in Africa, the campaign is on disinformation more generally. Similarly, TikTok has created a 'Be Informed' educational video series regarding media literacy, which is not focused on election disinformation.[222]

Google has created, in collaboration with U.S. based companies a literacy project to enhance online literacy among young people to help them recognize quality content as part of their policies surrounding election disinformation.[223]

b. Behaviour-based policies

221Facebook Newsroom 'Supporting Elections Across Africa'. About Facebook, 2020, https://about.fb.com/news/2020/10/supporting-elections-across-africa/. Accessed 15 Feb 2021.
222 Hind, Stephanie. Tiktok Newsroom, 2020, https://newsroom.tiktok.com/en-us/tiktoks-be-informed-series-stars-tiktok-creators-to-educate-users-about-media-literacy. Accessed 15 Feb 2021.
223 Gingras, Richard. 'Elevating Quality Journalism On The Open Web'. Google The Keyword, 2018, https://blog.google/outreach-initiatives/google-news-initiative/elevating-quality-journalism/. Accessed 15 Feb 2020.

Twitter seems to use the most behavioural type policies, encouraging users towards authoritative information and public discourse, instead of leaning on content removal. Twitter's civil integrity policy indicates to users that they may not post or share content that might 'suppress participation' in a civic process (such as an election) or mislead people about when, where, or how to participate.[224] Tweets with such content may be labelled as potentially misleading and contain a link to authoritative election information.[225]

Twitter may also provide election labels for certain elections (such as the 2020 U.S. election).[226] These labels appear on candidate's account pages and on every Tweet sent and Retweeted, and include the office they are running for, their state or district number identifier, and a small ballot box icon.[227] Additionally, as of August 2020, Twitter has labelled key government official's account and state-affiliated media accounts for the five permanent members of the UN Security Council, but with the option of expanding in the future.[228] According to Twitter this is to help provide users context and protect the public discourse between users and government leaders and officials.[229]

Twitter also announced a special policy in anticipation of the U.S. 2020 election, but which applied globally, called 'Quote Tweet'.[230] When users went to 'Retweet' another user's post, they were encouraged to give their own commentary instead of simply sharing the original post. Twitter explained that they hoped this policy would encourage users to add their own 'thoughts, reaction, and perspectives' instead of simply amplifying those of other users.[231] However, users had the option of declining and simply Retweeting. This policy ran only through the U.S. election and was not permanent.[232]

c. Restrictive policies

Twitter bans content that contains misleading information about how to participate in civic processes, voting suppression or intimidation, or misleading information about election outcomes.[233] Depending on the severity and type of the violation, Tweets may be directly deleted, or an account banned, although for less severe violations a warning or label is applied.[234]

YouTube has a similar policy banning content that includes voter suppression, false claims related to eligibility for political candidacy, or incitement to interfere with democratic processes (such as interrupting voting procedures).[235] When any such content is identified it is deleted (with no apparent option of labelling for less severe violations).[236] Facebook, which has a general policy of not removing content, has indicated

224 Twitter. 'Civic Integrity Policy'. Twitter Help Center, 2021, https://help.twitter.com/en/rules-and-policies/election-integrity-policy. Accessed 15 Feb 2021.
225 Ibid
226 Coyne, Bridget. 'Helping Identify 2020 US Election Candidates On Twitter'. Twitter Blog, 2019, https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html. Accessed 15 Feb 2021.
227 ibid
228 @TwitterSupport . 'New Labels For Government And State-Affiliated Media Accounts'. Twitter Blog, 2020, https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html. Accessed 15 Aug 2021.
229 Ibid
230 Gadde, Vijaya, and Kayvon Beykpour. 'Additional Steps We're Taking Ahead Of The 2020 US Election, (Updated 2 November 2020)'. Twitter Blog, 2020, https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html. Accessed 15 Feb 2021.
231 ibid
232 Ibid
233 Twitter. 'Civic Integrity Policy'. Twitter Help Center, 2021, https://help.twitter.com/en/rules-and-policies/election-integrity-policy. Accessed 15 Feb 2021.
234 Ibid
235 YouTube. 'Spam, Deceptive Practices, & Scams Policies'. Youtube Help, https://support.google.com/youtube/answer/2801973?hl=en#:~:text=YouTube%20doesn't%20allow%20spam,violates%20this%20policy%2C%20report%20it. Accessed 15 Feb 2021.
236 Ibid

that it might remove disinformation that would interfere with people voting (such as false information about dates and locations of voting).[237]

TikTok's U.S. election integrity policy allows for content removal, blocking accounts from future livestreaming, removing accounts and all account content, and banning accounts from the device, depending on the severity of the content infraction and the number of violations.[238]

Other policies focus more on de-amplifying such problematic content. Twitter again seems to have the most policy in this area. If a Tweet violates the civic engagement policy and is labelled, in additional to providing a warning to users and a link to alternative information, Twitter may de-activate comments on the Tweet, reduce its visibility, or prevent the Tweet from being recommended to other users.[239] Twitter used other de-amplification techniques during the U.S. 2020 election, when it prevented 'liked by' and 'followed by' recommendation appearing on user's timeline from other users you are not following.[240] Twitter directly acknowledged that this policy would likely reduce how often or quickly users see content from accounts they do not follow.[241]

Twitter also de-amplifies content by state-affiliated media accounts in additional to labelling them (only within the permanent members of the UN Security Council).[242] Facebook's policy is to not remove false news, choosing instead to de-amplify by showing any such content lower down on users' newsfeeds, thus slowing down its spread.[243] Facebook's rationale for such policy is that it does not want to 'stifle' public discourse, nor accidentally remove satire or parody.[244]

d. Political advertisements

How these companies handle political advertisements range from outright bans to measures aimed at increasing transparency in political advertising. This is a complex area when it comes to the protection of human rights and the freedom of expression. While it is important for political actors to be able to spread their messages so that voters can make informed choices, political advertisement can simultaneously spread disinformation about candidates, political parties, individuals or organizations.

Twitter[245] and TikTok[246] have banned political advertisements outright, while many of the other platforms choose to use behavioural measures instead. Reddit requires comments to be turned on for all political advertisements for at least the first twenty-four hours so that users can engage in discourse.[247] Reddit also lists information for all political ads in a specific 'sub reddit' to allow for transparency.[248] Snapchat needs a 'paid for by' banner on all political advertisements and does not allow any content that is misleading or

237 Facebook Newsroom 'Supporting Elections Across Africa'. About Facebook, 2020, https://about.fb.com/news/2020/10/supporting-elections-across-africa/. Accessed 15 Feb 2021.
238 Twitter. 'Civic Integrity Policy'. Twitter Help Center, 2021, https://help.twitter.com/en/rules-and-policies/election-integrity-policy. Accessed 15 Feb 2021.
239 Ibid
240 T Gadde, Vijaya, and Kayvon Beykpour. 'Additional Steps We're Taking Ahead Of The 2020 US Election, (Updated 2 November 2020)'. Twitter Blog, 2020, https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html. Accessed 15 Feb 2021.
241 Ibid
242 @TwitterSupport . 'New Labels For Government And State-Affiliated Media Accounts'. Twitter Blog, 2020, https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html. Accessed 15 Aug 2021.
243 Facebook. '21. False News'. Community Standards, https://www.facebook.com/communitystandards/false_news. Accessed 15 Feb 2021.
244 Ibid
245 Gadde, Vijaya, and Kayvon Beykpour. 'Additional Steps We're Taking Ahead Of The 2020 US Election, (Updated 2 November 2020)'. Twitter Blog, 2020, https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html. Accessed 15 Feb 2021.
246 'Integrity For The US Elections'. Tiktok Safety Center, https://www.tiktok.com/safety/resources/2020-us-elections?lang=en. Accessed 15 Feb 2021.
247 u/con_commenter. Reddit.Com R/Announcements, 2020, https://www.reddit.com/r/announcements/comments/g0s6tn/changes_to_reddits_political_ads_policy/. Accessed 15 Feb 2021.
248Ibid

deceptive in political ads.[249] Google similarly requires election advertisements for all States covered by Google election ad policies to include a 'paid for by' disclosure.[250]

II. COVID-19

Twitter defines COVID-19 related misleading content as: False or misleading information about the nature of the virus, the efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease, information about official regulations, information about the prevalence of the virus, or risk of infection or death.[251] This is an example of what kind of disinformation surrounds this pandemic. Most platforms such as YouTube won't allow content that spreads medical disinformation that contradicts local health authorities' or the World Health Organization's (WHO) medical information about COVID-19.[252] However generally most policies will not regulate strong commentary, opinions and/or satire, provided these do not contain false or misleading assertions of fact, direct responses to misleading information, personal anecdotes or first-person accounts and public debate about the advancement of COVID-19 science and research, such as effectiveness of research.[253]

Companies have been trying to limit the spread of such disinformation through a variety of policies. We'll now analyse some of the Covid-19 specific policies of some the major social media platforms.

Covid-19			
	Interactive policies	Behaviour policies	Restrictive policies
Facebook	☐	☑	☑
Twitter	☐	☑	☑
Reddit	☑	☐	☑
Youtube	☐	☑	☑
Google	☐	☐	☑
TikTok	☑	☑	☑
Amazon	☐	☐	☑
Snapchat	☑	☐	☑

a. Interactive policies

Companies have been using interactive policies that connect their users to experts in a bid to engage them and to help them wade through the waves of disinformation regarding this pandemic. Reddit has been organizing 'Ask Me Anything' (AMA) series in which users can ask scientific and medical professionals, as well as public officials, questions about the virus, enabling users to access verified, real-time information.[254]

249 Snapchat. 'Snap Political & Advocacy Advertising Policies'. Snap.Com, 2019, https://snap.com/en-US/ad-policies/political. Accessed 15 Feb 2021.
250 Google. 'Political Content'. Google Advertising Policies Help, 2021, https://support.google.com/adspolicy/answer/6014595?hl=en-GB&ref_topic=1626336. Accessed 15 Feb 2021.
251 Twitter. 'COVID-19 Misleading Information Policy.' Twitter Help Center, https:// help.twitter.com/en/rules-and policies/medical-misinformation-policy. Accessed 15 Feb. 2021.
252 YouTube. 'COVID-19 Medical Misinformation Policy -.' YouTube Help, https://support.google.com/youtube/answer/9891785?hl=en. Accessed 15 Feb. 2021.
253 Twitter. 'COVID-19 Misleading Information Policy.' Twitter Help Center, https:// help.twitter.com/en/rules-and-policies/medical-misinformation-policy. Accessed 15 Feb. 2021.
254 u/worstnerd. 'R/ModSupport - Misinformation and COVID-19: What Reddit Is Doing.' Reddit, 2020, www.reddit.com/r/ModSupport/comments/g21ub7/disinformation_and_covid19_what_reddit_is_doing. Accessed 15 Feb. 2021.

TikTok has launched two hashtags[255] with the aim to involve the community with educational live programs as well as providing entertainment to encourage users to stay at home and slow the spread of the virus.[256] In October the app has also announced another hashtag that a group of scientists and clinicians will use to share stories about their daily work and educate users on vaccines.[257]

In an attempt to limit spread of disinformation through their platform, SnapChat created the 'friend check-up' feature which will prompt users to look at their friend list and ensure that they only keep close connections to avoid connections with strangers that may spread disinformation.[258] The platform has also created stickers and filters that connect users directly with the WHO that users can add to their photos and videos that promote correct information on how to prevent the spread of the virus.[259]

b. Behavioural policies

Companies have been using an array of behavioural polices to contain the spread of disinformation regarding the COVID-19 pandemic. Twitter has applied a label on Tweets containing potentially harmful and misleading information related to the pandemic.[260] Depending on the propensity for harm and type of misleading information, warnings may also be applied. These warnings will inform people that the information in the Tweet is in conflict with public health experts' guidance before they can view it.[261] For content related to COVID-19 to be labelled by Twitter it must: advance a claim of fact, expressed in definitive terms; be demonstrably false or misleading, based on widely available, authoritative sources; and be likely to impact public safety or cause serious harm.[262]

Facebook instead sends messages to those who interacted with disinformation about COVID-19 that has been removed. Users will receive a notification that states that the platform has removed a post they've interacted with for violating the policy regarding COVID-19 disinformation that leads to imminent physical harm. Once they click on the notification, they will see a thumbnail of the mentioned post, and they will be able to access information regarding where they saw it and how they engaged with it. They will also be able to see why it is false and why it was removed. Finally, users will then be able to see more facts about COVID-19 in Facebook's Coronavirus Information Center and take other actions such as unfollowing the Page or Groups that shared the misleading content. [263] The company is also improving search results on both Facebook and Instagram. When looking up 'vaccine' or 'COVID-19' the search promotes relevant, authoritative results and provide third-party resources to connect people to expert information[264]

TikTok instead will redirect searches associated with vaccine or COVID-19 disinformation to their Community guidelines and will not autocomplete anti-vaccine hashtags in the search bar.[265] Furthermore, users who choose to explore hashtags related to the virus will be met with an in-app notice that provides

255 #HappyAtHome and #DistnaceDance, TikTok. 'Safety Center - Resources | TikTok.' TikTok Safety Center, www.tiktok.com/safety/resources/covid-19?lang=en. Accessed 15 Feb. 2021.
256 TikTok. 'Safety Center - Resources | TikTok.' TikTok Safety Center, www.tiktok.com/safety/resources/covid-19?lang=en. Accessed 15 Feb. 2021.
257 TikTok. '#TeamHalo Shines a Light on TikTok.' Newsroom | TikTok, 2020, newsroom.tiktok.com/en-gb/teamhalo-shines-a-light-on-tiktok. Accessed 15 Feb. 2021.
258 Fischer, Sara. 'Snapchat Urges Users to Remove Unwanted Connections.' Axios, 2021, www.axios.com/snapchat-unwanted-connections-c495029b-8900-42e8-938d-c68cd51152cc.html. Accessed 15 Feb. 2021.
259 Ibid
260 Twitter. 'COVID-19 Misleading Information Policy.' Twitter Help Center, help.twitter.com/en/rules-and-policies/medical-misinformation-policy. Accessed 15 Feb. 2021.
261 Ibid
262 Ibid
263 Rosen, Guy. 'An Update on Our Work to Keep People Informed and Limit Misinformation about COVID-19, Update on December 15, 2020.' About Facebook, 2020, about.fb.com/news/2020/04/covid-19-misinfo-update/. Accessed 15 Feb. 2021.
264 Rosen, Guy. 'An Update on Our Work to Keep People Informed and Limit Misinformation about COVID-19, Update on February 8, 2021.' About Facebook, 2020, about.fb.com/news/2020/04/covid-19-misinfo-update/. Accessed 15 Feb. 2021.
265 'Safety Center - Resources | TikTok.' TikTok Safety Center, www.tiktok.com/safety/resources/covid-19?lang=en. Accessed 15 Feb. 2021.

direct access to WHO's website and other local health agencies. This is accompanied by a reminder to report content that violates community standards.[266]

Finally, other platforms such as YouTube have simply put up a panel with links to national health agencies[267]

c. Restrictive policies

As we have just mentioned companies try to limit the impact of the contact of users with disinformation through labels and banners, once these are applied the content can be de-amplified or removed.

For example, once Twitter applies the labels, that have been mention in the previous section, the tweets won't be amplified and will have a more limited reach. These labels are applied though machine learning, which may lack context, therefore the platform will not permanently suspend accounts that spread disinformation based solely on an automated enforcement system. They will however require users to remove tweets that include denial of health authority recommendations, description of alleged cures, description of harmful or ineffective treatments or measures and claims that intend to manipulate people into certain behaviour for the gain of a third party or cause panic. Furthermore, the platform will require the elimination of content that includes claims by people impersonating a government or health official, Propagating false or misleading information around the virus diagnostic criteria and procedures. The platform adopts a similar approach to vaccines.[268]

Twitter's advertisement policies allow for the advertisement of content having implicit or explicit reference to COVID-19 only if it refers to: Adjustments to business practices and/or models in response to COVID-19, Support for customers and employees related to COVID-19, restrictions may apply[269] Twitter has also prohibited the creation of fake accounts which misrepresent their affiliation or share content that falsely claims affiliation to a medical practitioner, public health official or agency, research institution, or that falsely suggests expertise to speak informatively COVID-19 related issues unless they fall under parody, newsfeed, commentary, or fan accounts.[270]

In December Facebook announced that it would start removing claims about Covid-19 and its vaccines that have been debunked by public health experts. Such claims are prohibited in advertisements as well.[271] Following an investigation about the spreading of Covid-19 related disinformation Facebook removed the capabilities to target people that are interested in 'pseudoscience' which may be more vulnerable to misleading claims.[272]

Regarding sales of cures, Amazon has prohibited the sale of products that claim to cure, mitigate, treat, or prevent diseases in humans or animals without FDA (Food and Drug Administration) approval (including COVID-19). Products that claim to be 'FDA-Cleared,' 'FDA-Approved' or products that include the FDA logo in associated images need to meet added requirements. Finally, selling products that are marketed with

---

266 Ibid

267 Rasool, Aqsa. 'How YouTube Is Dealing with Misleading Coronavirus Videos?' Digital Information World, 2020, www.digitalinformationworld.com/2020/04/how-youtube-is-dealing-with-misleading-coronavirus-videos.html. Accessed 15 Feb. 2021.

268 The platform may label, place a warning or remove Tweets that advance unsubstantiated rumours, disputed claims, as well as incomplete or out-of-context information about vaccines. Twitter. 'COVID-19 Misleading Information Policy.' Twitter Help Center, help.twitter.com/en/rules-and-policies/medical-misinformation-policy. Accessed 15 Feb. 2021.

269 such as Distasteful references to COVID-19 (or variations) are prohibited, Content may not be sensational or likely to incite panic, Prices of products related to COVID-19 may not be inflated, the promotion of certain products related to COVID-19 may be prohibited. Titter. 'Inappropriate Content.' Twitter Business, business.twitter.com/en/help/ads-policies/ads-content-policies/inappropriate-content.html. Accessed 15 Feb. 2021.

270 Twitter. 'COVID-19 Misleading Information Policy.' Twitter Help Center, help.twitter.com/en/rules-and-policies/medical-misinformation-policy. Accessed 15 Feb. 2021.

271 Rosen, Guy. 'An Update on Our Work to Keep People Informed and Limit Misinformation about COVID-19, Update onFebruary 8, 2021.' About Facebook, 2020, about.fb.com/news/2020/04/covid-19-misinfo-update/. Accessed 15 Feb. 2021.

272 Sankin, Aaron. 'Want to Find a Misinformed Public? Facebook's Already Done It – the Markup.' Themarkup.org, 2020, themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it. Accessed 15 Feb. 2021.

environmental claims must ensure that they are not misleading about the qualities or characteristics of a product.[273]

TikTok prohibits false and misleading content, including that related to Covid-19 and vaccines this information, such content will be removed. the company does not allow paid advertising that advocates against vaccination although PSAs related to COVID-19 vaccines are accepted on a case-by-case basis as long as they are in the interest of public health and safety.[274]

Reddit uses its moderators to remove misleading content, obvious disinformation will be deleted by automated rules and other kinds of disinformation need to be reported by users.[275]

Finally, Google News does not publish medical content from any site that contradicts or runs contrary to scientific or medical consensus and evidence-based best practices.[276]

III. Manipulated media

Manipulated media, including the more recently developed deep fakes, are one of the techniques used to spread disinformation through the modification of media such as videos or photos in order to lead others to believe in harmful or untrue narratives. With the recent development of AIs and deep learning makes it more difficult to recognize media that has been manipulated. [277] With these new developments Tech companies have created policies that try to limit the harm done by manipulated media.

Manipulated media			
	Interactive policies	Behaviour policies	Restrictive policies
Facebook	☐	☐	☑
Twitter	☐	☑	☑
Reddit	☐	☐	☐
Youtube	☐	☐	☑
Google	☐	☐	☑
TikTok	☐	☑	☑
Amazon	☐	☐	☐
Snapchat	☐	☐	☑

a. Behavioural policies

273 Amazon United States . 'Prohibited Product Claims.' Amazon Seller Central, sellercentral.amazon.com/gp/help/external/G202024200language=en_US#:~:text=The%20Food%20and%20Drug%20Administration. Accessed 15 Feb. 2021.
274 TikTok, 'Supporting Our Community Through COVID-19', TikTok Safety Center https://www.tiktok.com/safety/resources/covid-19?lang=en Accessed 15 Feb. 2021.
275 u/worstnerd. 'R/ModSupport - Misinformation and COVID-19: What Reddit Is Doing.' Reddit, 2020, www.reddit.com/r/ModSupport/comments/g21ub7/disinformation_and_covid19_what_reddit_is_doing. Accessed 15 Feb. 2021.
276 Google. 'Google News Content Policies - Publisher Center Help.' Support.google.com, support.google.com/news/publisher-center/answer/6204050?visit_id=637486411155795195-1321351808&rd=1. Accessed 15 Feb. 2021.
277 Sample, Ian. 'What Are Deepfakes – and How Can You Spot Them?' The Guardian, The Guardian, 13 Jan. 2020, www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

Twitter has implemented a cumulative set of criteria to determine whether to label or remove manipulated media.[278] This is a test based on whether the subject appears to be synthetic or manipulated (in the form of additional visual and auditory information and depictions of persons are obviously simulated), whether the subject is shared in a deceptive manner, and whether the content is likely to impact public safety or cause serious harm (i.e., if the user creating the content shows dangerous patterns like stalking or obsessive attention, or if the subject is likely to provoke civil unrest).[279] If the content is manipulated, then, Twitter may attach a warning before the option to 'like' or 'retweet' or attach a banner to the publication showing that the information had been manipulated and supply alternative sources of information.[280]

b. Restrictive policies

Content must reach a certain level of threat to the public order to warrant its removal. Facebook has policies which trigger the removal of content if it is obvious that it has been synthesized to manipulate truth or fact. Facebook has a policy that manipulation will be removed if ' ... it has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say.'[281] Products of artificial intelligence and machine learning that superimposes content onto a video, fabricating authenticity, warrants its removal from Facebook.[282]

If content appears to be synthesized to manipulate users, it will be subjected by Facebook to review by independent fact-checkers.[283] If rated false or partially false, distribution in users' News Feed will be reduced and advertisements will be rejected. Warning signs alerting falsity will be shown to people who have shared and seen the post.[284] The Deep Fake Detection Challenge (DFDC) was launched to accelerate development of new ways to detect deep fake videos.[285]

Twitter, likewise, removes tweets which are manifestly fabricated to manipulate truth.[286] Both Snapchat[287] and TikTok[288] may also remove manipulated media which is misleading. Furthermore, YouTube policies prevent posting of manipulated media which is classified as information which has been manipulated in a misleading manner and may pose a serious risk of egregious harm.[289]

As for Google, when it comes to manipulated media, audio and visual content (videos and imagery) that has been edited to deceive, fraud or mislead by means of fabricating actions or events that verifiably did not take place that poses a high risk of fundamental misunderstanding, must be taken down and prevented from being broadcasted.[290] The standard at which this must be assessed is the likelihood of causing significant harm to groups or individuals, or significantly undermining participation or trust in electoral or civic processes. Additionally, Google play does not allow apps (applications) which enable users to

278 Twitter. 'Building Rules in Public: Our Approach to Synthetic & Manipulated Media.' Twitter.com, 2020, blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html. Accessed 15 Feb. 2021.
279 Ibid
280 Ibid
281 Bikert, Monika. 'Enforcing against Manipulated Media.' About Facebook, 6 Jan. 2020, about.fb.com/news/2020/01/enforcing-against-manipulated-media/.
282 Ibid
283 Ibid
284 Ibid
285 Google. 'Google News Content Policies - Publisher Center Help.' Support.google.com, support.google.com/news/publisher-center/answer/6204050?visit_id=637486411155795195-1321351808&rd=1. Accessed 15 Feb. 2021.
286 Twitter. 'Building Rules in Public: Our Approach to Synthetic & Manipulated Media.' Twitter.com, 2020, blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html. Accessed 15 Feb. 2021.
287 Snapchat. 'Community Guidelines - Snap Inc.' Snap.com, snap.com/en-US/community-guidelines. Accessed 15 Feb. 2021.
288TikTok, 'Community Guidelines.' Www.tiktok.com, www.tiktok.com/community-guidelines?lang=en. Accessed 15 Feb. 2021.
289 ---. 'Spam, Deceptive Practices, & Scams Policies - YouTube Help.' Support.google.com, support.google.com/youtube/answer/2801973?hl=en#:~:text=YouTube%20doesn. Accessed 15 Feb. 2021.
290 Google, 'Google News content policies ', news publisher help forum, https://support.google.com/news/publisher-center/answer/6204050?visit_id=637486411155795195-1321351808&rd=1 Accessed 15 Feb. 2021.

distribute misleading information in the form of manipulated media clips or false text messages.[291] Apps which generate manipulated content without transparency safeguards or promoting misleading claims will be removed.[292]

IV. Impersonation and Identity fraud

Coupled with identity fraud, impersonation on the internet causes multiples types of harm. This includes harm to individual users, who may see their own image or page impersonated, societal harms when users lose trust in one another or their news sources, as well as potential security harms, such as might happen if out-of-State profiles impersonate and spread information within another State. Prompted by the anonymity that users enjoy in the online world, the most common mode of online impersonation is either by stealing one's information to gain access to his online profile or by creating a completely fake profile.[293] To limit the harm brought by such practices the various companies have created policies and taken measure that are mostly restrictive.

Impersonation Identity fraud			
	Interactive policies	Behaviour policies	Restrictive policies
Facebook	☐	☑	☑
Twitter	☐	☑	☑
Reddit	☐	☐	☐
Youtube	☐	☐	☑
Google	☐	☑	☑
TikTok	☐	☐	☑
Amazon	☐	☐	☐
Snapchat	☐	☐	☑

a. Behavioural policies

In June 2009 Twitter introduced account verification[294] with a distinct blue tick badge next to the name of individuals or organisations. This was closely followed by Google in 2011,[295] Facebook in 2012[296] and Instagram in 2014.[297] For Twitter, a verified badge is applied if an account is authentic, notable and active. The criteria 'notable' is defined as association with a prominently recognised individual or brand, broadly categorised under 'government', 'companies, brands and organisations', 'news organisations and

291 Google, 'Google Play is helping to safeguard elections', blog post, 19 August 2020,https://blog.google/outreach-initiatives/civics/google-play-helping-safeguard-elections/ Accessed 15 Feb. 2021.
292 Google, 'Google Play is helping to safeguard elections', blog post, 19 August 2020,https://blog.google/outreach-initiatives/civics/google-play-helping-safeguard-elections/ Accessed 15 Feb. 2021.
293 Kambellari, Evisa. 'Online Impersonation: I Have a Right to Be Left Alone v. You Can't Mandate How I Use My Privacy Toolbox.' Ssrn.com, 2017, papers.ssrn.com/sol3/papers.cfm?abstract_id=3351507.
294 Cashmore, Pete. 'Twitter Launches Verified Accounts.' Mashable, mashable.com/2009/06/11/twitter-verified-accounts-2/?europe=true. Accessed 15 Feb. 2021.
295 Mead, Derek. 'Google+ Now Verifying Accounts of the Famous | Digital Trends.' Digital Trends, 21 Aug. 2011, www.digitaltrends.com/social-media/google-now-verifying-accounts-of-the-famous/. Accessed 15 Feb. 2021.
296 J Constine, Josh. 'Facebook Launches Verified Accounts and Pseudonyms.' TechCrunch, techcrunch.com/2012/02/15/facebook-verified-accounts-alternate-names/. Accessed 15 Feb. 2021.
297 D'Onfro, Jillian. 'Instagram Is Introducing 'Verified Badges' for Public Figures.' Business Insider Nederland, www.businessinsider.nl/author/jillian-donfro?international=true&r=US. Accessed 15 Feb. 2021.

journalists', 'entertainment', 'sports and gaming', and 'activists, organizers, and other influential individuals'.[298]

b. Restrictive policies

For identity fraud, such as users posing as someone they're not or deceiving people about their identity, most of the platforms solely use restrictive policies in preventing such behaviour. Many of such restrictions include removal of the profile, account or page, along with all content. Snapchat,[299] and Instagram,[300]for example, listed such prohibition under their general terms of use.

For Twitter, notwithstanding the allowance to create parody, newsfeed, commentary, or fan accounts, impersonation is a violation in which accounts that pose as another person, brand, or organization in a confusing or deceptive manner may be permanently suspended.[301] However, an account will not be removed if the user shares merely share the same name but has no other commonalities, or the profile clearly says it is not affiliated with or connected to any similarly named individuals or brands.[302] TikTok's identity fraud policy is drafted similarly: when reports of impersonation are made, the platform will request for user review or ban the account.[303] Amazon disallows any content or listing which uses, misuses, portrays affiliation with, or otherwise uses another party's brand in a manner that is confusing to customers.[304]

Facebook does not allow the use of their services for impersonation and will disable any account that impersonates others, defined as using others' photos with the explicit aim to deceive, creating an account assuming to be or speak for another person or entity or creating a page assuming to be or speak for another person or entity for whom the user is not authorized to do so.[305] For cases of misrepresented account, compromised accounts and empty account with prolonged dormancy, verification and further information will be requested before actions such as temporary restriction or permanent disable of account are taken.[306]

YouTube drafted their identity protection policy to allow for complaints if someone's video or personal information is posted without their consent.[307] YouTube will then ask the uploader to remove, and should an agreement not be reached, request can be made to YouTube for the removal of content based on their privacy guidelines.[308] However, the impersonation policy is stricter, wherein content intended to impersonate a person and channel is strictly prohibited on the platform.[309] Channel impersonation is defined as channels which copy another channel's profile, background, or overall look and feel in such a way that makes it look like someone else's channel.[310] The channel does not have to be 100% identical if the intent is clear to copy the other channel. Personal impersonation, on the other hand are content intended to look like someone else is posting it.[311] Furthermore, channels, or content in the channel, which causes confusion about the source of goods and services advertised are not allowed on the platform as well.[312]

---

298 Twitter. 'About Verified Accounts.' Help.twitter.com, help.twitter.com/en/managing-your-account/about-twitter-verified-accounts#:~:text=The%20blue%20verified%20badge%20on. Accessed 15 Feb. 2021.
299 Snapchat. 'Community Guidelines - Snap Inc.' Snap.com, snap.com/en-US/community-guidelines. Accessed 15 Feb. 2021.
300 'Terms of Use | Instagram Help Center.' Instagram.com, 2015, help.instagram.com/581066165581870. Accessed 15 Feb. 2021.
301 Twitter Help Center, 'Impersonation policy', https://help.twitter.com/en/rules-and-policies/twitter-impersonation-policy Accessed 15 Feb. 2021.
302Ibid
303 'TikTok Community Guidlines.' Www.tiktok.com, www.tiktok.com/community-guidelines?lang=en#37. Accessed 15 Feb. 2021.
304 'Amazon.com Help: Content Guidelines for Books.' Amazon.com, 2021, www.amazon.com/gp/help/customer/display.html?nodeId=15015801&language=en_US&ref=efph_home_cont_200164670. Accessed 15 Feb. 2021.
305 Facebook community standards, 'Account integrity and Authentic identity', https://www.facebook.com/communitystandards/misrepresentation Accessed 15 Feb. 2021.
306 Ibid
307 YouTube Help, 'Protecting your identity', https://support.google.com/youtube/answer/2801895?hl=en Accessed 15 Feb. 2021.
308 Ibid
309 YouTube Help, 'Policy on impersonation', https://support.google.com/youtube/answer/2801947 Accessed 15 Feb. 2021.
310 Ibid
311Ibid
312 Ibid

With a general rule that dates and by-lines, as well as information about authors, the publication, the publisher, company or network must be clearly visible, Google News also has strict impersonation policy.[313] Not only sites or accounts that impersonate any person or organisation are banned, sites or accounts that misrepresent or conceal their ownership or primary purpose are disallowed too. Inauthentic or coordinated behaviour that misleads users are prohibited, such as misrepresentation or concealment their country of origin, and/or sites or accounts working together in ways that conceal or misrepresent information about their relationships or editorial independence, and/or content that conceals or misrepresents sponsored content as independent, editorial content.[314] Furthermore, sponsorship, including, but not limited to, ownership or affiliate interest, payment or material support, should be clearly disclosed to readers. The subject of sponsored content should not focus on the sponsor without clear disclosure. [315]

V. Fake engagement

Fake engagement means artificially influencing how well rated or viewed your content is on a platform. This can take the form of false ratings, increasing your number of 'followers' using bots, or otherwise inflating your account or content's visibility. Such fake engagement may increase the visibility of misleading or false information, which has ramifications for the right to access information, as it may become more difficult for users to understand which information is correct and which is not. Content or accounts that seem as though they have many 'followers' or high ratings may appear more credible to users.

Fake engagement can also increase the spread of potentially harmful extremism, conspiracies and hate speech. For example, if an individual user has an opinion that is extremist and seemingly unpopular (such as that women owe men sex, as can be found in so called 'Incel' forums), he may understand that his opinion is fringe and not generally supported by most platform users. If, however, that user then views content espousing the same opinion which has a falsely inflated number of views or ratings he may come to believe that his opinion is shared and commonly accepted. This can help to spread and solidify such extremist opinions and can ultimately lead to harm to others.

Companies can work to reduce such false inflation in ways that do not negatively impact users' freedom of expression. For example, to limit the spread of disinformation, YouTube has changed its recommendation system. This system was previously based on how many 'clicks' a video received.[316] YouTube found that such a system can contribute to misleading information as content creators try to create 'click bait' to get views. YouTube has changed their algorithms to consider how long a user watched a video, requiring a certain amount of watch time before the video received another 'view' on its counter.[317]

313 Google Publisher Center Help, 'Google News content policy', https://support.google.com/news/publisher-center/answer/6204050?visit_id=637486411155795195-1321351808&rd=1 Accessed 15 Feb. 2021.
314 Ibid
315 Ibid
316 How Google Fights Disinformation. , 2019. P.20 Available at https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf?hl=en Accessed 15 Feb. 2021
317 Ibid

Fake Engagement			
	Interactive policies	Behaviour policies	Restrictive policies
Facebook	☐	☐	☑
Twitter	☐	☐	☑
Reddit	☐	☐	☐
Youtube	☐	☐	☑
Google	☐	☐	☐
TikTok	☐	☐	☑
Amazon	☐	☐	☑
Snapchat	☐	☐	☐

a. Restrictive policies

Companies seem to rely solely on restrictive policies when combating fake engagement and metric inflation, opting to remove problematic content and ban users. This response is understandable, as it is conceivably difficult to create warnings or labels which accurately and simply describe to users where the potential danger may lie when they are viewing content which has had its metrics falsely inflated.

TikTok's fake engagement policy indicates that the company will remove content or accounts that violate the policy.[318] Prohibited behaviour includes sharing instructions on how to artificially increase views, likes, followers, shares or comments; attempting to or engaging in selling or buying views, etc.; promoting artificial traffic generation services; or operating multiple accounts under false pretences in order to distribute commercial spam.[319] Similarly, YouTube's policies prohibit the artificial increase of views, likes, or comments and any content which exists solely to incentivize viewers for engagement.[320] Prohibited actions include linking to services which would artificially inflate metrics, offering to subscribe to other users' channels in return for their subscribing to yours, or creating any content which features other users buying views.[321] Any content which violates this policy is removed.[322]

Twitter also has a similar policy, prohibiting metric inflation, selling or purchasing followers or engagement, reciprocal inflation, or selling or trading accounts.[323] Twitter relies on restrictive measures for policy violation, which could take the form of Tweet deletion, locked account, or permanent suspension.[324]

Amazon also relies on restrictive policies for violations of its ratings, feedback, and reviews policy. This policy prohibits sellers from paying for or offering incentives for customers to provide or remove feedback or product reviews; asking customers to only write positive reviews or to remove or change reviews;

318 'Community Guidelines.' Www.tiktok.com, www.tiktok.com/community-guidelines?lang=en. Accessed 15 Feb. 2021.
319 Ibid
320 YouTube Help Center, 'Fake Engagement Policy',
https://support.google.com/YouTube/answer/3399767?hl=en&ref_topic=9282365 Accessed 15 Feb. 2021.
321 Ibid
322YouTube Help Center, 'Fake Engagement Policy',
https://support.google.com/YouTube/answer/3399767?hl=en&ref_topic=9282365 Accessed 15 Feb. 2021.
323 Twitter Help Center, 'Platform manipulation and spam policy', September 2020, https://help.twitter.com/en/rules-and-policies/platform-manipulation#:~:text=You%20may%20not%20use%20Twitter's,disrupts%20people's%20experience%20on%20Twitter.&text=To%20make%20that%20possible%2C%20we,other%20types%20of%20platform%20manipulation. Accessed 15 Feb. 2021.
324Ibid

soliciting reviews from customers who had positive experiences; or sellers reviewing their own products or a competitor's products.[325]

3(b). To What Extent Do You Find These Measures To Be Fair, Transparent And Effective In Protecting Human Rights, Particularly Freedom Of Opinion And Expression?

I. Transparency

While it is to be commended that these platforms have all created policies trying to diminish the harm of misleading and false information on their platforms, it is of the upmost importance that such policies are transparent to users. Transparency allows users to understand what type of content they may not post, why their content may have been removed or de-amplified, and why they may be seeing or not seeing certain content. Lack of transparency could have a chilling effect on freedom of speech if users lose trust in these platforms.

The most transparent policies clearly define what types of content users are prohibited from posting, together with examples of such content. Twitter's system of clearly defining terms such as 'civic processes'[326] is such an example, as well as YouTube's thorough use of examples of the categories of content which is prohibited.[327]

Transparency is also required as to what actions may be taken in response to the posting of banned content and how decisions are made. Twitter[328] and YouTube's[329] use of a clear three-strike rule are transparent, as users can easily find not just what content is prohibited, but what actions may be taken and what that means for their account's standing. Users can foresee that after they receive three strikes, their account may be deactivated. Facebook's creation of an Oversight Board is to be applauded as a step toward transparency and fairness.[330]

Amazon's policies lack such transparency in decision making. While there are policies on certain content that Marketplace postings cannot have,[331] their selling policy states, 'Amazon reserves the right to determine the appropriateness of listings on its site and remove any listing at any time.'[332] Similarly, Amazon's content policies for entertainment media lack transparency, such as including under content rules for books: 'We reserve the right to remove content from sale if we determine it creates a poor customer experience', without clarity on what a 'poor customer experience' might mean.[333] In March of 2019 Amazon pulled anti-vaccination movies and books from its platforms after it received criticism, without officially announcing

---

325 Amazon seller central, 'Selling Policies and Seller Code of Conduct' https://sellercentral.amazon.com/gp/help/external/G1801?language=en_US Accessed 15 Feb. 2021.

326 Twitter Help Center, 'Civic Integrity Policy', January 20201, https://help.twitter.com/en/rules-and-policies/election-integrity-policy Accessed 15 Feb. 2021.

327 See for example, YouTube Help Center, 'Spam, deceptive practices, & scams policy', https://support.google.com/YouTube/answer/2801973?hl=en&ref_topic=9282365 Accessed 15 Feb. 2021.

256 Twitter Help Center, 'Civic integrity policy', https://help.twitter.com/en/rules-and-policies/election-integrity-policy Accessed 15 Feb. 2021.

329 YouTube Help Center, 'Community Guidelines strikes basics', https://support.google.com/YouTube/answer/2802032 Accessed 15 Feb. 2021.

330 Oversight Board webpage, oversight board.com; more about this Oversight Board can be found under Question 4.

331 Amazon, seller central, 'Selling Policies and Selling Code of Conduct', https://sellercentral.amazon.com/gp/help/external/G1801?language=en_US Accessed 15 Feb. 2021.

332Amazon seller central, 'Offensive and Controversial Materials', https://sellercentral.amazon.com/gp/help/external/200164670 Accessed 15 Feb. 2021.

333 Amazon, Help & Customer Service, 'Content Guidelines for Books', https://www.amazon.com/gp/help/customer/display.html?nodeId=15015801&language=en_US&ref=efph_home_cont_200164670 Accessed 15 Feb. 2021.

this move or answering questions as to why or how it decided what media to remove.[334] This is unacceptable from a framework of the right to freedom of expression and access to information. While a company may be justified in removing content that has misleading health information, there must be transparency in this process. *Platforms should not remove content without notifying users or the public and offering explanations based on pre-existing policies.*

The move towards transparency in advertising on these sites, both political advertising and otherwise, is welcomed. This can take the form of labels and banners placed on advertisements showing who the purchaser is, as Google has done in many countries for political advertisements.[335] Facebook's[336] and Reddit's[337] use of advertisement 'libraries' is also a step in the right direction, as they allow users to easily find information on advertisements users are viewing.

One area in which these companies need to work toward improving transparency is the use of algorithms for suggesting content and for flagging content that potentially violates the terms of service. Google explains their use of such algorithms, 'Ranking algorithms are an important tool in our fight against disinformation. Ranking elevates the relevant information that our algorithms determine is the most authoritative and trustworthy above information that may be less reliable.'[338] Users need more transparency in how algorithms are taught about what information is or is not reliable and to what extent their content or other content may be recommended to others. However, this transparency becomes even more important when algorithms are used to find content to flag or remove.

To fully ensure transparency, these companies should regularly publish reports outlining how much and what types of content are subject to behavioural or restrictive measures. Currently, Reddit's transparency report does not outline how much content is being removed by admins or mods under the platform's disinformation policies.[339] The transparency reports of both Facebook and Twitter only contain general statistics about request for content removal but they do not disclose what is removed and why[340] thus defeating the purpose of transparency.

Facebook provides access to an 'Ads Library' where users can find information on all advertisements currently running on any Facebook app or Instagram (owned by Facebook).[341] The aim of this program is to achieve more transparency for political and social issue advertisements. For example, users can look up how much certain pages, such as political candidates' pages, have spent on advertising and which advertisements they purchased.[342] Facebook has recently attempted to open this information to researchers through its Facebook Open Research & Transparency (FORT)[343] platform, which allows researchers to download data on how Facebook advertisements are targeted to users. Reddit has a similar subreddit

334 Business, Jon Sarlin, CNN. 'Anti-Vaccine Movies Disappear from Amazon after CNN Business Report.' CNN, edition.cnn.com/2019/03/01/tech/amazon-anti-vaccine-movies-schiff/index.html. Accessed 15 Feb. 2021.
335 Google, advertising policy help, 'Political content', https://support.google.com/adspolicy/answer/6014595?hl=en-GB&ref_topic=1626336 Accessed 15 Feb. 2021.
336 Leathern, Rob, Facebook newsroom, 'Expanded Transparency and More Controls for Political ads', 9 January 2020, https://about.fb.com/news/2020/01/political-ads/ Accessed 15 Feb. 2021.
337 u/con_commenter. Reddit.Com R/Announcements, 2020, https://www.reddit.com/r/announcements/comments/g0s6tn/changes_to_reddits_political_ads_policy/. Accessed 15 Feb 2021.
338 How Google Fights Disinformation. , 2019. P12 Available at https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf?hl=en Accessed 15 Feb 2021.
339 'How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19.' New America, www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/reddit/. Accessed 15 Feb. 2021.
340 Wade, M., Walsh, M. J., & Baker, S. A. (n.d.). Disinformation: tech companies are removing 'harmful' coronavirus content – but who decides what that means? The Conversation, 27 August 2020, https://theconversation.com/disinformation-tech-companies-are-removing-harmful-coronavirus-content-but-who-decides-what-that-means-144534 Accessed 15 Feb. 2021.
341 Facebook Business Help Center, 'About the ad library', https://www.facebook.com/business/help/2405092116183307?id=288762101909005 Accessed 15 Feb. 2021.
342 Ibid
343 Facebook, 'Increasing Transparency Around US 2020 Election Ads', 25 January 2021, https://about.fb.com/news/2021/01/increasing-transparency-around-us-2020-elections-ads/ Accessed 15 Feb. 2021.

dedicated to transparency in political advertisements. Twitter has created an archive of Tweets which it believes were a part of state-backed information operations to influence political elections, so that users can see past Tweets that may have contained misleading or false information.[344]

II. Effectiveness

Digital technology companies' policies on misleading and false information need to be effective in protecting freedom of expression, as well as other human rights, such as the rights to non-discrimination and personal security. This is a balancing act, as protecting freedom of expression to an extreme can result in harm to some individuals or groups (such as when false information about a minority group leads to greater discrimination). Additionally, the right to freedom of expression also has a right to access to information. This right can be both hindered and protected by limitations to the freedom of expression. If users are not able to adequately share information and ideas, this limits access to information; however, if users are able to share any misleading or false information they want, other users may have trouble finding correct information. Because of these inherent conflicts between and within rights, **we consider as most effective policies which take harm to others into account when restricting misleading or false information or which otherwise prevent individual or societal harms,** as these types of policies most effectively and efficiently protect all rights at stake.

a. Policies which take harm to others into account:

For example, Twitter's policy on manipulated media includes three criteria for what content may be labelled or removed:[345]

- Are the media synthetic or manipulated?
- Are the media shared in a deceptive manner?
- *Is the content likely to affect public safety or cause serious harm?* - threats to a specific person, risks of mass violence or civic unrest, threats to privacy or ability to freely participate

Twitter provides this graphic to help users understand when their content is likely to be labelled or removed. As can be seen, content is only likely to be removed when it is 'likely to impact public safety or cause serious harm'. The policy further clarifies that harm includes threat to the privacy of an individual or group, or the ability of an individual or group from freely expressing themselves or taking part in civil events.[346] This includes stalking, obsessive attention, targeted content that aims to silence someone, or voter suppression or intimidation.[347]

344 Twitter, Information Operations' report, https://transparency.twitter.com/en/reports/information-operations.html Accessed 15 Feb. 2021.

345 Twitter, Help Center, 'Synthetic and manipulated media policy', https://help.twitter.com/en/rules-and-policies/manipulated-media Accessed 15 Feb. 2021.

346 Ibid

347 Ibid

Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled.
✗	✓	✗	Content **may** be labeled.
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

Twitter, Help Center, 'Synthetic and manipulated media policy', https://help.twitter.com/en/rules-and-policies/manipulated-media

Similarly, YouTube bans and will remove any manipulated content that misleads users and may pose a serious risk of egregious harm.[348]

Twitter's policy on COVID-19 misleading information also takes harm into account, showing that depending on the propensity for harm and type of misleading information, a warning may be applied to misleading content in addition to an ordinary label.[349] Such a warning would inform users that the information in the Tweet conflicts with public health experts' guidance before they view it. The COVID-19 policy further clarifies that for a tweet to be labelled or removed due to having misleading information, three criteria must be met:[350]

- advance a claim of fact, expressed in definitive terms.
- be demonstrably false or misleading, based on widely available, authoritative sources; and
- *be likely to affect public safety or cause serious harm.*

Such a policy effectively protects against harm that may be caused by misleading information around the virus, while also allowing for freedom of expression. In fact, the policy specifically notes that strong commentary, opinions, satire, counter-speech, personal anecdotes, or public debate about the advancement of COVID-19 science and research are not violations of this policy.[351] Thus, Twitter strikes a fair balance between protecting freedom of expression and protecting from potential harms caused by expression.

b. Policies which prevent individual or societal harms:

The above policies show that more than potential physical harm must be considered. The concept of harm must cover non-physical harms to specific groups, such as minorities, as well as harms to society. Considerations of the latter can be seen in the above COVID-19 policy of Twitter, which considers that harm is caused where content misleads individuals about the causes, prevention, or treatment of the virus.[352] This type of harm can also arise in the context of disinformation about vaccines. Facebook has expanded its COVID-19 misleading information policy to prohibit and remove false claims about COVID-19 and its vaccines which have been debunked.[353] On the other hand, Amazon's policies do not seem to consider the harm that can be caused by anti-vaccine disinformation. A recent study on how Amazon's algorithms

348 YouTube Help Center, 'Spam, deceptive practices, & scams policy', https://support.google.com/YouTube/answer/2801973?hl=en&ref_topic=9282365 Accessed 15 Feb. 2021.
349 Twitter Help Center, 'COVID-19 misleading information policy', https://help.twitter.com/en/rules-and-policies/medical-disinformation-policy Accessed 15 Feb. 2021.
350 Ibid
351 Ibid
352 Ibid
353 Rosen, Guy. 'An Update on Our Work to Keep People Informed and Limit Misinformation about COVID-19, Update onFebruary 8 2021.' About Facebook, 2020, about.fb.com/news/2020/04/covid-19-misinfo-update/. Accessed 15 Feb. 2021.

suggest anti-vaccine related products found that a higher content of vaccine disinformation comes up when making vaccine related searches, and these are suggested over 'debunking' content.[354]

Harm to minority groups must be taken into account as well, such as in Facebook and TikTok's Holocaust denial policies, and TikTok's policy to remove content that contains false information about minorities. However, it is not enough for these companies' policy to facially take harm against minorities into account. They must ensure that their policies and automated content detection methods can adequately protect against false and misleading information directed against minority communities. During the Black Lives Matter protests of the last few years, and especially during the protests during the summer of 2020, many companies, including Facebook, seemed unable to effectively restrict false information aimed at increasing racial conflict and discrimination against Black communities and users.[355] This issue of policies causing harm to minorities will be covered more extensively below under 'Fairness'.

Similarly, these companies must ensure that their policies have the capacity to reduce harm in order to be effective. Facebook's general policy is to not remove false information.[356] But such a policy, which seems to err on the side of leaving material up, can become problematic if the policies in place are not effective in reducing harm. The Facebook algorithm may be helping amplify information around the pandemic by boosting content that receives a lot of engagement. Sensational content, such as health disinformation, can receive significant engagement and as a consequence might be boosted by the algorithm, thus giving a greater platform to disinformation above authoritative sources. This should have been curtailed by the labelling of misleading information as we have seen in the previous section however a 2020 Avaaz report found that Facebook failed to catch 84% of health-related disinformation content.[357] Google also has a policy of not removing content, but instead uses algorithms to 'elevate authoritative high-quality information' and providing tools to give users more contexts for their searches.[358] Such policies need to be critically analysed to determine whether they can truly offer users, as well as protected groups, protection from harm.

III. Fairness

There are multiple dimensions to the concept of fair protection of human rights, especially when competing rights are at stake, as they are within the realm of disinformation. Fairness includes transparency, as it is unfair to expect users to follow rules which they cannot understand or foresee. Fairness also includes access to a grievance or appeal policy, as well as monitoring of the use of company policies, all of which will be covered in more depth below in answer to question 3(c). We will focus our attention on two other dimensions of fairness. The first is how well policies allow for expressions which may have some misleading or false information, but which are not harmful. This is like the balancing of rights as described under 'Effectiveness', but concerns situations when there is no foreseeable harm in expressions, such as satire and parody. Second, fairness requires that policies dealing with disinformation must equally protect all users.

---

354 Juneja, Prerna, and Tanushree Mitra. 'Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation.' ArXiv:2101.08419 [Cs], 29 Jan. 2021, arxiv.org/abs/2101.08419, 10.1145/3411764.3445250. Accessed 15 Feb. 2021.; See also Wiggers, Kyle. 'University of Washington Researchers Say Amazon's Algorithms Spread Vaccine Misinformation.' VentureBeat, 26 Jan. 2021, venturebeat.com/2021/01/26/university-of-washington-researchers-say-amazons-algorithms-spread-vaccine-misinformation/. Accessed 15 Feb. 2021.

355 'Anti-Racism Protests: Divisive Disinformation Narratives Go Viral on Facebook, Racking up over 26 Million Estimated Views.' Avaaz, 2020, secure.avaaz.org/campaign/en/anti_protest_disinformation/ see also ; Dazed. 'TikTok Says Issues with Black Lives Matter Hashtags Were due to a Bug.' Dazed, 30 May 2020, www.dazeddigital.com/science-tech/article/49419/1/tiktok-says-issues-with-black-lives-matter-george-floyd-hashtags-due-to-a-bug. Accessed 15 Feb. 2021.

356'Community Standards | Facebook.' Www.facebook.com, www.facebook.com/communitystandards/false_news. Accessed 15 Feb. 2021.

357 Section 1.2-1.3 'Facebook's Algorithm: A Major Threat to Public Health.' Avaaz, 2020, secure.avaaz.org/campaign/en/facebook_threat_health/. Accessed 15 Feb. 2021.

358 How Google Fights Disinformation., 2019. P.10 Available at https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf?hl=en Accessed 15 Feb. 2021.

First, disinformation policies need to allow for expressions that facially have false or contested information, but which do not do harm. For example, satire and parody policies are necessary to the protection of freedom of expression. Such policies allow users to engage with false information in a way that does not harm others. Twitter's misleading information policies make exceptions for satire and parody if the content does not include false information in a misleading manner.[359] Similarly, content which critically considers disputed information is necessary for the public discourse of ideas and information. YouTube's policies allow content that violates their disinformation policies but that either gives weight to countervailing views from authorities, or if the purpose is to condemn or dispute the disinformation.[360] This policy allows for content that is not likely to cause harm, but which grapple with disputed information. TikTok's general Community Guidelines allow for exceptions to all their policies, including their disinformation policy, for 'educational, documentary, scientific, or artistic content, satirical content, content in fictional settings, counter speech, and content in the public interest that is newsworthy or otherwise enables individual expression on topics of social importance.'[361] To protect these non-harmful expressions, companies also need to ensure that any use of automated content flagging or censorship does not include this content in its net.

Second, disinformation policies must actually protect all users. As described under the 'Effectiveness' section, overly protecting freedom of expression can potentially minority groups by allowing the spread of 'discriminatory disinformation'. However, restrictions on freedom of expression can also cause harm to minority groups, if such groups are unfairly targeted by these restrictions.

Instagram has been accused of prioritizing white content creators and supporting companies which only use thin, white influencers to promote their product, contributing to the normalization of mediocre representation of people of colour. Such practices take a heavy toll on young, impressionable users. A lack of representation of POC puts whiteness a pedestal and keeps it as an unattainable norm. The Head of Instagram, Adam Mosseri, issued a statement in June 2020, acknowledging a racial bias in the platform. He said that '(b)lack people are often harassed, afraid of being 'shadow banned', and disagree with many content takedowns.'[362]

A four-pronged solution was introduced to review their policies to end racial bias: an examination of online harassment (in the form of hate speech), account verification, distribution (i.e., filtering people without transparency and making them less likely to show up on the 'Explore' feature for exposure), and algorithmic biases which 'repeat patterns developed by our biased societies.'[363] However, there has not yet been any update or publication of the review. Whether these reviews of bias and proposed reforms will be effective in discouraging racially charged practices has yet to be seen.

Facebook has also had problems with their disinformation policies actually causing harm to minority groups. Multiple accounts have surfaced of Black users' content on racism being flagged as hate speech and removed.[364] Facebook has responded, saying that is attempting to re-configure its algorithms to better protect Black users and allow for constructive content about racism.[365]

This is a difficult area, as often it cannot be proven that company's algorithms may be disproportionately affecting minority users, and the companies themselves may not always be aware of how their policies and content review systems may be unintentionally or indirectly biased against certain groups. It is generally

359 Twitter Help Center, 'COVID-19 misleading information policy', https://help.twitter.com/en/rules-and-policies/medical-disinformation-policy Accessed 15 Feb. 2021.
360 YouTube Help, 'COVID-19 Medical Disinformation Policy',
https://support.google.com/YouTube/answer/9891785?hl=en&hl=en&ref_topic=9282436 Accessed 15 Feb. 2021.
361 Tik Tok, 'Community Guidelines', https://www.tiktok.com/community-guidelines?lang=en#37 Accessed 15 Feb. 2021.
362 'Ensuring Black Voices Are Heard | Instagram Blog.' About.instagram.com,
about.instagram.com/blog/announcements/ensuring-black-voices-are-heard. Accessed 13 Feb. 2021.
363 Ibid
364 Guynn, Jessica. 'Facebook While Black: Users Call It Getting 'Zucked,' Say Talking about Racism Is Censored as Hate Speech.' USA TODAY, eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/. Accessed 15 Feb. 2021.
365 Dwoskin, Elizabeth et. al. 'Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show'. The Washington Post. 3 December 2020. https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/. Accessed 15 Feb. 2021.

the individual users who notice such patterns. Transparency and accountability become of the upmost importance. As suggested above, companies should regularly publish data on how much and what types of content are subject to behavioural and restrictive measures, as well as make public information about the use of their algorithms. And perhaps more importantly, *these companies need to take seriously any user allegations of discrimination caused by their policies and automated content systems*.

When users feel that they are being discriminated against on a platform and believe that their content is being unfairly censored, this causes individual harm. But it also results in societal harm: we need minority voices to be a part of the public discourse, and we need to be able to trust that we have access to those voices. Therefore, harm occurs when there are allegations of such indirect discrimination which companies do nothing to address or disprove.

3(C). What Procedures Exist To Address Grievances And Provide Remedies For Users, Monitor The Action Of The Companies, And How Effective Are They?

All of the digital technology companies considered in this report have some form of appeal or grievance mechanism by which users penalized for violating disinformation policies can seek remedy. However, these processes are dealt with internally and the companies' decisions are not externally monitored. They also often lack transparency as to decision making. Thus, their effectiveness in protecting users and freedom of speech is often limited.

YouTube's appeal policy is one of the most thorough. Users can appeal either a strike received on their account or appeal a video removal. YouTube is clear about what might happen as a result of an appeal:

- If we find that your content followed our Community Guidelines, we'll reinstate it and remove the strike from your channel. If you appeal a warning and the appeal is granted, the next offense will be a warning.
- If we find your content followed our Community Guidelines, but isn't appropriate for all audiences, we'll apply an age-restriction. If it's a video, it won't be visible to users who are signed out, are under 18 years of age, or have Restricted Mode turned on.
- If we find that your content was in violation of our Community Guidelines, the strike will stay, and the video will remain down from the site. There's no additional penalty for appeals that are rejected.
- You may appeal each strike only once.[366]

To contrast, Twitter's appeal option only seems to apply to a suspension or locked account, and it is not clear whether a lesser penalty can be appealed, such as a label, de-amplification, or a strike.[367] Similarly, Amazon has an appeals process for deactivated accounts and listing removals, but it is not clear whether there is such a process for other penalties.[368] Google's appeal process being limited to rejected advertisements[369] makes sense in the overall context of Google's policy not to remove content.

These type of limited or unclear appeals processes are not effective at protecting the rights of users, as they lack transparency and accountability. Instead, these vague policies seem to be aimed at keeping power over content moderation completely within the company. For example, Snapchat's Community Guidelines

366 YouTube. 'Appeal Community Guidelines Actions'. YouTube Help. https://support.google.com/YouTube/answer/185111?hl=en&ref_topic=9387060. Accessed 15 Feb. 2021.

367 Twitter. 'Appeal an account suspension or locked account'. Help Center. https://help.twitter.com/forms/general. Accessed 15 Feb. 2021

368 Amazon. 'Appeal an account deactivation or listing removal'. Seller Central. https://sellercentral.amazon.com/gp/help/external/G200370560?language=en_US. Accessed 15 Feb. 2021.

369 Google. 'Fix a disapproved ad'. Ads Help. https://support.google.com/google-ads/answer/1704381?hl=en-GB. Accessed 15 Feb. 2021.

states, 'we reserve the right to decide, in our sole discretion, what content violates that spirit and will not be permitted on the platform.'[370]

But as described in the preamble to this report, Governments are no longer willing to allow companies this complete control over content regulation. In order for these companies to balance protecting the rights of users and maintaining an internal system of oversight, they need to create appeals and grievance process which are more transparent, and which include an element of external monitoring. Facebook's Oversight Board is an example of such a policy.

Facebook has set up an Oversight Board whose purpose is to promote free expression through independent decision-making regarding Facebook and Instagram's content. This Board is staffed by independent members from diverse backgrounds and disciplines who are appointed by a trust.[371] The Board selects cases that are difficult, significant and globally relevant, thus not all user appeals will be considered by the Board.[372] After Facebooks makes a first decision on user content, such as to remove the content, the company will let the user know if the decision is eligible for appeal to the Board.[373] However, no guidelines have been published listing what content may be eligible for appeal. [374]

4. Please Share Information On Measures That You Believe Have Been Especially Effective To Protect The Right To Freedom Of Opinion And Expression While Addressing Disinformation On Social Media Platforms.

In general terms it could be said that the most effective measures are those that aim at developing transparency, independent oversight, the right to appeal, access to effective remedies and democratic control over removal of content. The challenge with any measure tackling disinformation is balancing the need to prevent the spread of potentially harmful information, with the need to guarantee and respect fundamental rights, most notably freedom of expression (creation of content).

One innovative measure for combating disinformation while providing safeguards for the freedom of expression is the **Facebook Oversight Board ('FOB')**, an attempt by Facebook to create an external private quasi-judicial body to provide oversight of content moderation (especially content removal). The FOB defines its purpose as 'the promotion of free expression by making principled, independent decisions regarding content on Facebook and Instagram and by issuing recommendations on the relevant Facebook company content policy'[375].

The Board gives users more ability to be heard than do other such appeals processes.[376] This increased ability to be heard provides two concrete safeguards of users' rights. The first is that it makes Facebook more accountable for its policies by providing transparency into how decisions on content removal are

370 Snap Chat. 'Community Guidelines' https://snap.com/en-US/community-guidelines. Accessed 15 Feb. 2021.
371 Oversight Board. 'Ensuring Respect for Free Expression, Through Independent Judgement'. https://oversightboard.com/. Accessed 15 Feb. 2021.
372 Oversight Board. 'Appealing Content Decisions on Facebook or Instagram'. Appeals Process. https://oversightboard.com/appeals-process/. Accessed 15 Feb. 2021.
373 Oversight Board. 'Appealing Content Decisions on Facebook or Instagram'. Appeals Process. https://oversightboard.com/appeals-process/. Accessed 15 Feb. 2021.
374 Facebook. 'How do I appeal Facebook's content decision to the Oversight Board?'. Facebook Help Center. 2021. https://www.facebook.com/help/346366453115924. Accessed 15 Feb. 2021.
375 'Oversight Board | Independent Judgement. Transparency. Legitimacy.'. Oversightboard.Com, https://oversightboard.com/. Accessed 11 Feb 2021.
376 K. Klonick, The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. Yale Law Journal 129, no. 8 (2418-2499), at 2491 (2020).

made. Secondly, it provides the user an effective remedy against content moderation exercised by Facebook and Instagram.

Another productive solution to balance the right to freedom of expression while limiting disinformation is digital technology company's use of external **fact-checking,** as well as **collaboration with external experts.** External fact-checking policies protect users, as they remove some of the control over content regulation from the digital technology companies themselves. Use of external experts helps users find correct and authoritative information rather than just limiting disinformation. When companies utilise fact-checkers, it is important that these groups follow the principles outlined by the International Fact Checking Network.[377]

Other effective measures **include users in how disinformation is prevented**. For example, LinkedIn empowers users with the option to Disable, Re-enable, and Limit Comments on Posts.[378] In this way, while users can still engage with contents by liking and sharing, while disinformation can be controlled by users.

Twitter has launched a pilot measure to be implemented first for its US user, called **Birdwatch** which will also allow for user participation in the limiting of disinformation. The new policy will allow users to interact with tweets that they consider to be misleading by including a note that they consider relevant to give context to that information[379]. This tool can empower users and stimulate engagement.

5. Please Share Information On Measures To Address Disinformation That You Believe Have Aggravated Or Led To Human Rights Violations, In Particular The Right To Freedom Of Opinion And Expression.

Online hinderances to freedom of expression

Automated systems can be an efficient means of detecting contested materials posted online, but they could also lead to the violations of the freedom of expression, as they are not always able to effectively contextualize words, sentences and expressions.[380] For instance, Facebook labelled a passage from the Declaration of Independence as hate speech and removed the content, because it was not programmed to put posts into context.[381] This is only one example of many instances when automated systems accidentally removed information. When automated detection systems can remove content without certification by a human, it can lead to patterns which hinder the individual's right to freedom of expression.

Another means of protecting people from disinformation is by banning users from social media platforms when users have a history of disseminating disinformation or inciting discrimination or violence. The most famous example is Twitter banning Donald Trump after he incited an insurrection on Capitol Hill in January 2021.[382] Banning users is an effective preventative measure but it could also lead to discrimination and violation of the freedom of expression. Although platforms need to have this option in their "back

377 'International Fact-Checking Network - Poynter'. Poynter, 2021, https://www.poynter.org/ifcn/. Accessed 11 Feb 2021.
378 From the LinkedIn option: 'Comments have been turned off on this post. You can still reactor share it.'
379 Coleman, Keith. 'Introducing Birdwatch, a community-based approach to misinformation.' Twitter Blog. 25 Jan. 2021. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html. Accessed 15 Feb. 2021.
380 Aswad, E.M. 'The future of freedom of expression online'. Duke Law & Technology Review, Vol. 17, 2018.
381Annie Grayer, CNN. 'Facebook Labels Part Of Declaration Of Independence 'Hate Speech"'. CNN, 2021, https://www.cnn.com/2018/07/05/politics/facebook-post-hate-speech-delete-declaration-of-independence-mistake/index.html. Accessed 15 Feb 2021.
382 Twitter. 'Permanent Suspension of @realDonaldTrump'. Twitter Blog. 8 Jan. 2021. https://blog.twitter.com/en_us/topics/company/2020/suspension.html. Accessed 15 Feb. 2021.

pocket",[383] disinformation policies should also contain options for less restrictive measures, such as labelling, de-amplification, or content removal.

Furthermore, a common occurrence which hinders rights of users is a lack of transparency and corrective oversight. This problem was highlighted in *Google v. Spain,*[384] where an arbitrary body was created within Google to deal with requests for taking down content. This can be problematic because of the obscurity of the practises, without independent oversight or means of appeal, can lead to unfair censorship.[385]

A worrying hindrance to the freedom of expression is government-sanctioned internet shutdowns and firewalls. These measures are an extreme form of control over free information and serious violations of freedom of opinion and thought. Authoritarian governments use this tactic under the guise of limiting disinformation, with intention to control media and sources of information available to the general public. For example, Indian government officials at local and national levels have often said that the many Internet shutdowns that have taken place were necessary to keep law and order and prevent the spread of disinformation.[386]

In a similar limb to firewalls and blocking websites, the creation of alternative sites with heavy governmental influence is another method of controlling the spread of information and regulating censorship. Local versions of social media can be a tool to actualising this goal. Examples of this practise are seen in Egypt,[387] and in China.[388]

Censorship during the Covid-19 Pandemic

A few European governments, especially in eastern and central Europe, have used the ongoing health crisis as a pretext to restrict the free flow of information and clamp down on independent media.[389] For example, the Hungarian government passed a law that criminalises the spreading of 'false' or 'distorted' information that undermines preventing the spread of COVID-19. Sanctions for which are fines and up to five years imprisonment. Journalists and media freedom advocates fear the interpretation of these new measures will be weaponised to silence what remains of the country's independent press.[390]

On March 16, the president of Romania, Klaus Iohannis, signed an emergency decree which gives authorities the power to remove reports or take down websites that spread "fake news" about the virus. The measure lacks any means of availing of an effective remedy.[391] In Bulgaria, the government used the state of emergency decree to try to amend the penal code and introduce prison sentences for spreading 'fake news' about Covid-19. This measure included sanctions of up to three years in prison or a fine of up to €5,000.[392] Russia demanded more than twenty media outlets remove content it deemed 'inaccurate,

383 See for example: 'Facebook, Apple, Youtube And Spotify Ban Infowars' Alex Jones'. The Guardian, 2021,https://www.theguardian.com/technology/2018/aug/06/apple-removes-podcasts-infowars-alex-jones. Accessed 15 Feb 2021. ; Aswad, E.M. 'The future of freedom of expression online'. Duke Law & Technology Review, Vol. 17, 2018.

384 CJEU. Case C- 131/12, Google Spain, ECLI:EU:C:2014:317

385 Leiser, M. 'Private jurisprudence' and the right to be forgotten balancing test.' Computer Law & Security Review, Volume 39, (2020) https://www.sciencedirect.com/science/article/pii/S0267364920300637. Accessed 15 Feb. 2021.

386(www.dw.com), Deutsche. 'India's Internet Shutdowns Function Like 'Invisibility Cloaks' | DW | 13.11.2020'. DW.COM, 2021, https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554. Accessed 15 Feb 2021.

387 Loc.Gov, 2019, https://www.loc.gov/law/help/social-media-disinformation/social-media-disinformation.pdf. Accessed 15 Feb 2021.

388 Loc.Gov, 2019, https://www.loc.gov/law/help/social-media-disinformation/social-media-disinformation.pdf. Accessed 15 Feb 2021.

389 'Fighting Europe's Covid-19 Infodemic: Freedom Of Expression Concerns'. Csis.Org, 2021, https://www.csis.org/blogs/technology-policy-blog/fighting-europes-covid-19-infodemic-freedom-expression-concerns. Accessed 10 Feb 2021.

390'European Media Freedom Suffers Under COVID-19 Response - International Press Institute'. International Press Institute, 2020, https://ipi.media/european-media-freedom-suffers-covid-19-response. Accessed 11 Feb 2021.

391 'Manufacturing Censorship (Consent Not Required)'. CIJ, 2020, https://cji.ro/en/manufacturing-censorship-consent-not-required/. Accessed 11 Feb 2021.

392'European Media Freedom Suffers Under COVID-19 Response - International Press Institute'. International Press Institute, 2020, https://ipi.media/european-media-freedom-suffers-covid-19-response. Accessed 11 Feb 2021.

socially significant information' about the coronavirus from their websites.[393] While the spread of disinformation is crucial in a global pandemic, legally based safeguards must be implemented to prevent abuse of power and censorship.

The link between hate speech and misinformation

Both the 2016 US election and the COVID-19 pandemic sparked debate about the role of social media in perpetuating discriminatory patterns embedded in our society. In a study undertaken by Muller and Schwarz, a link between hate crimes and the use of social media was established.[394] In their findings, it was established that areas with a high percentage of Facebook users also had high levels of racially based hate crimes and anti-immigrant sentiment. This pattern was especially demonstrated in places with high racial and ethnic tensions. The spread of disinformation with underlying racial tones is incredibly dangerous because of the amount of people who believe fake news at face value and can exacerbate racial biases and hatred. An example of the use of disinformation within hate speech is anti-Asian sentiments created by the conspiracy theory that coronavirus is a Chinese 'invention' purposefully spread to undermine the West, which led to increased discrimination and acts of hate against Asian communities in Europe and in the United States.

Muller and Schwarz established links between disinformation and discrimination.[395] Social media algorithms link people who genuinely believe these conspiracies and find themselves in online spaces where this narrative is not only normalised but encouraged. These online spaces foster isolated ideologies and normalise fringe communities. [396] This allows harmful ideologies (such as white supremacy and misogyny) to thrive. Reinforcement of extreme ideologies and challenge avoidance are more likely to happen on social media. Interestingly, studies have shown that social media usage is responsible for spreading extreme political ideology and disinformation more than the news. This is because people are less likely to stop watching the news if concepts under discussion contrast one's political ideology or orientation.[397] Furthermore, social media provides the tools to switch to content which aligns with one's own beliefs.

However, despite inherent flexibility to switch to more 'agreeable' content, this can result in a social media 'filter bubble.'[398] This is the use of personalised and/or automated algorithms to determine how social media platforms rank information on a person's feed. This reduces civic engagement and the extent to which people hear from the other side of the argument. Unsurprisingly, racially charged opinions are left unchallenged. The spread of algorithmic racially based discrimination adds to political polarization; for example, anti-immigration policies are attractive to right-wing conservatives who believe in narrow application of immigration laws. In an interview with Larry King, Barack Obama spoke out about the effect the 'filter bubble' and algorithms have on politics:

> "*If you are getting all your information off algorithms being sent through your phone and it's just reinforcing whatever biases you have, which is the pattern that develops, at a certain point you just live in a bubble, and that's part of why your politics is so polarised right now.*"[399]

Also see 'Council Of Europe Says Media Freedom In Bulgaria On Decline'. Seenews.Com, 2020, https://seenews.com/news/council-of-europe-says-media-freedom-in-bulgaria-on-decline-693329. Accessed 11 Feb 2021.

393 http://www.washingtontimes.com, The. 'Russia Invokes 'Fake News' Law To Order Removal Of Coronavirus Reports From Web'. The Washington Times, 2020, https://www.washingtontimes.com/news/2020/mar/20/russia-invokes-fake-news-law-to-order-removal-of-c. Accessed 11 Feb 2021.

394 Muller and Schwartz, 'Fanning the Flames of Hate: Social Media and Hate Crime' 2020.

395 Muller and Schwartz, 'Fanning the Flames of Hate: Social Media and Hate Crime' 2020.

396 Pablo Barbera, 'Social Media, Echo Chambers and Political Polarisation', Cambridge University Press, 2020.

397 Pablo Barbera, 'Social Media, Echo Chambers and Political Polarisation.

398 Pablo Barbera, 'Social Media, Echo Chambers and Political Polarisation', Cambridge University Press (2020) pg. 41.

399 Pablo, Barbera. 'Social Media, Echo Chambers and Political Polarisation'. Cambridge University Press (2020) pg. 35.

According to Reppell and Shein,[400] disinformation campaigns which use hate speech as a tactic relies on underlying social dynamics and existing divisive messages. According to their model of the dissemination of hate speech and its amplification on social media, there are five steps.[401] First, the producers of the speech come up with an ideologically motivated expression based on racial bias or their own belief or world view. Second, the message is spread promoting intolerance and violence against groups by direct or indirect reference to race, nationality, ethnicity, religion, gender, disability or sexual orientation. Thirdly, hate speech can go viral through social media and amplified without any restrictive or preventative measures in place on the site. Fourthly, the message is interpreted by likeminded people. Finally, the risk of hate speech dissemination may lead to undermining faith in democratic institutions, democratic participation, as well as an increase of hate crimes and citizen polarisation.

When nefarious actors target false or misleading information at a certain group or community this can result in further stigmatisation and violence against that group. When hate speech policies, either Governmental policies or policies of digital technology companies, do not address false information, this leads to reduced protections of human rights, such as the right to access to information and the right to non-discrimination. In conclusion, the prevention of dissemination of disinformation can undermine freedom of expression if used in an authoritarian manner, demonstrated by governments abovementioned. This practise, paired with disinformation based on xenophobic ideology, puts human rights at an enormous risk.

6. Please Share Any Suggestions Or Recommendation You May Have For The Special Rapporteur On How To Protect And Promote The Right To Freedom Of Opinion And Expression While Addressing Disinformation.

- **Transparency of Political Ads:** We recommend that the Special Rapporteur encourage Governments to regulate how digital technology companies ensure transparency of political advertisements. Examples to highlight include Australia's mandate for all paid electoral advertising, including advertisements on social media, to be authorized and to have an authorization statement.[402] Canadian law requires online platforms to keep and support a digital registry of all regulated ads related to federal elections, showing the names of agents who authorized them and any partisan advertising and election advertising that was published on the platform during election periods.[403] The European Commission issued a recommendation ahead of the 2019 European Parliament elections calling on European Union Member States to promote active disclosure of who is behind paid, online political advertisements and communications during electoral campaigns.[404] Transparency should also include transparency about the recipient of any micro-targeted advertisement, so that the marketplace of ideas and rebuttal can flourish.
- **Address connection between hate speech and disinformation:** When hate speech policies are considered, the role of disinformation is often ignored, and vice versa. But as described above under Question 4, false or misleading information targeted at a minority group can result in discrimination, stigmatization, and even violence against that group. Therefore, we suggest that the Special Rapporteur: include in her such 'discriminatory disinformation'; encourage all Governmental hate speech policies to include 'discriminatory disinformation'; and encourage all digital technology companies to create and enforce policies which prohibit false or misleading

400 'Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions,' 2019.
401 'Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions,' 2019, pg. 4.
402 Bradshaw, Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project, Oxford Internet Institute, 6;
403 ibid
404 ibid

information regarding any race, gender, religion, or other protected characteristic, and include such policies within COVID-19 specific disinformation policies and adopt policies.[405]

- **Effective Remedies:** In addressing the connection between hate speech and disinformation, we recommend the Special Rapporteur to reiterate the importance of the six-part criteria set out in the Rabat Plan of Action for criminalization of certain expressions. Furthermore, it should be highlighted that States should ensure that individuals who has suffered actual harm, should have access to effective remedies. This means that the affected individual should have access to an independent and impartial tribunal established by law.[406] Additionally, this Action Plan should also be used as a blueprint for digital technology companies to provide effective remedies for individuals who have suffered actual harm.[407] At the same time, the Special Rapporteur should remind States and companies to provide non-judicial remedies as well. Such remedies could include educational efforts on the adverse effects of disinformation and hate speech.
- **Independent Oversight** is crucial to supporting democratic ideals and principles. Organizations that are instrumental in decisions about the removal of content should subject themselves to either independent oversight or to independent audit by a body of human rights experts.

405 For example, Twitter will remove Tweets which claim that specific groups or nationalities are more susceptible to COVID-19 or are never susceptible to COVID-19. Twitter also bans any false or misleading information about the nature of the virus.
Facebook has created a new policy as of January 2021 to connect people with authoritative information about the Holocaust. Anyone searching for the terms associated with either the Holocaust or Holocaust denial, will see a message from Facebook encouraging them to connect with credible information about the Holocaust off Facebook. This has been done to curb anti-Semitism globally and decrease alarming level of ignorance about the Holocaust, especially among young people TikTok blocks searches for 'Holocaust denial' and other related terms, and directs users searching for Holocaust related terms to verified and authoritative information.
TikTok has also announced that it will not allow content that denies the Holocaust or other violent tragedies, and will remove such content, including content that contains disinformation about Jewish, Muslim, and 'other communities'. Additionally, TikTok will remove content that harms the LBGTQ+ community, including content that spreads the idea that individuals are not born LGBTQ+ or promotes conversion therapy.
406 Report Of The United Nations High Commissioner For Human Rights On The Expert Workshops On The Prohibition Of Incitement To National, Racial Or Religious Hatred. 2013, https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf. Accessed 15 Feb 2021. para 31
407Report Of The United Nations High Commissioner For Human Rights On The Expert Workshops On The Prohibition Of Incitement To National, Racial Or Religious Hatred. 2013, https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf. Accessed 15 Feb 2021.