UN Special Rapporteur for Freedom of Expression

Submission on an Annual Thematic Report on Disinformation

March 2021

Centre for Law and Democracy
info@law-democracy.org
+1 902 431-3688
www.law-democracy.org

Introduction[1]

This Submission provides an overview of actions proposed and taken by both governments and private sector actors (specifically, leading intermediaries) in response to disinformation. It assesses the impacts of these proposals and actions, primarily in light of their potential impact on the exercise of freedom of expression, as guaranteed under international law, such as in Article 19 of the *International Covenant on Civil and Political Rights*.[2]

While this Submission focuses on disinformation, meaning the intentional dissemination of inaccurate or misleading information, it also addresses misinformation, or the unintentional dissemination of such information. Third party acts of misinformation often massively boost the scale and speed of dissemination of disinformation, especially over social media networks, which also has a tendency to cleanse or launder the disinformation. As a result, there is a close relationship between the two issues, even if they are very different social phenomena, particularly in terms of the challenges in addressing disinformation.

The first part reviews various government responses to disinformation, ending up with a summary of better practice approaches and recommendations. The second part examines private sector responses, focusing on a few key players, and then summarises ongoing challenges and makes recommendations.

1. Government Responses to Disinformation

1.1 Government Anti-Disinformation Initiatives

While disinformation is often framed as an issue driven by private actors, public officials and other State actors can be a source of disinformation, and it is important that any response to disinformation stresses State obligations to refrain from spreading disinformation and instead to provide accurate information to the public. A survey by the International Center for Journalists of 1,406 journalists and media workers in 125 countries early in the COVID-19 pandemic asked respondents to identify their main sources of disinformation. While the largest response was for

[2] UN General Assembly Resolution 2200A (XXI), 16 December 1966, in force 23 March 1976, https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

"regular citizens" at 49%, 46% of respondents listed "political leaders and elected officials", 34% listed heavily partisan or State media, 25% named identifiable government agencies or their spokespeople, and 23% referenced government-sponsored troll networks.[3]

These numbers suggest that the starting place for governments seeking to combat disinformation should be inward-looking. A few initiatives have tried to do this. For example, in Uruguay, the Uruguayan Press Association supported an initiative to have the major political parties sign an Ethical Pact by which they committed not to generate disinformation during the upcoming elections.[4]

Generally, however, government anti-disinformation initiatives target external rather than internal sources of disinformation. Unfortunately, some government initiatives which are ostensibly aimed at combating dis- and misinformation have been blatantly propagandistic and at times themselves disseminated disinformation. For example, the Myanmar military established a "Tatmadaw True News Information Team" which released doctored or mislabelled photos related to the Rohingya crisis. Facebook eventually blocked the accounts of this team for spreading hate speech.[5] In Thailand, the Ministry of Digital Economy and Society's "Anti-Fake News Center", which aims to target news that undermines "peace and order, good morals and national security", works jointly with police in a manner that raises serious concerns about its impact on freedom of expression.[6]

Even government fact-checking initiatives that are less overtly propagandistic or opportunistic raise serious questions about independence and integrity. When such units are contained in government information offices or services, for example, they are not well-positioned to act as an independent fact-checking voice, although it is good for such offices to have internal protocols to ensure they are not disseminating disinformation.

Mexico, for example, generated significant controversy for establishing a "Verificado" fact-checking unit within the government's official news service, Notimex, which has the same name

---

[3] Julie Posetti, Emily Bell and Pete Brown, *Journalism and the Pandemic: A Global Snapshot of Impacts* (2020, ICJF and Tow Center for Digital Journalism), p. 14, https://www.icfj.org/sites/default/files/2020-10/Journalism%20and%20the%20Pandemic%20Project%20Report%201%202020_FINAL.pdf.

[4] Silvia Higuera, "On the Initiative of Journalists' Association, Political Parties in Uruguay to Sign a Pact against Misinformation", 24 April 2019, Knight Center LatAm Journalism Review, https://latamjournalismreview.org/articles/on-the-initiative-of-journalists-association-political-parties-in-uruguay-to-sign-a-pact-against-misinformation/.

[5] Poppy McPherson, "Exclusive: Fake Photos in Myanmar Army's 'True News' Book on the Rohingya Crisis", 28 December 2018, Reuters, https://www.reuters.com/article/us-myanmar-rohingya-photos-exclusive-idUSKCN1LF2LB; and Radio Free Asia, "Critics Warn of Deception as Myanmar Military Returns to Facebook After 2018", 10 June 2020, https://www.rfa.org/english/news/myanmar/military-facebook-06102020170558.html.

[6] Patpicha Tanakasempipat, "Thailand Unveils 'Anti-Fake News' Center to Police the Internet", 1 November 2019, Reuters, https://www.reuters.com/article/us-thailand-fakenews-idUSKBN1XB48O; and OHCHR, "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.

as an independent fact-checking source which fact-checked statements by the President. Questions were also raised about the lack of a clear methodology for fact checking.[7] India's Press Information Bureau established a fact-checking unit in 2019 (along with a COVID-19 section after the pandemic started) which reportedly initially focused on news about the government,[8] raising concerns about such a unit serving primarily to defend the government's reputation. Little information has been made available about how decisions labelling information as false are made. Pakistan's Ministry of Information and Broadcasting similarly has a "Fake News Buster" twitter account.[9]

A related issue is where such initiatives lack clarity in their operating rules and powers. For example, the Argentine Journalism Forum and other media groups objected strongly to the creation of a new observatory for detecting disinformation in Argentina, known as NODIO by its Spanish acronym. The entity was created within the Public Defender for Audiovisual Media, which receives complaints from the public about broadcast media.[10] Although NODIO did not have any formal sanctioning power, media actors were concerned that it would effectively operate as a State censorship tool, given the lack of clear legal standards or transparency about its role, structure or decision-making processes.[11]

A better approach is to offer funding or other support to independent fact-checking initiatives. South Africa, for example, has directed the public to "Real 411", a website run by Media Monitoring Africa, an independent non-governmental entity.[12]

---

[7] Syndy García, "Una Metodología Transparente es Indispensable en una Unidad de Verificación: Clara Jiménez", 22 July 2019, Verificado, https://verificado.com.mx/una-metodologia-transparente-es-indispensable-en-una-unidad-de-verificacion-clara-jimenez/; Vanguardia, "Acusan a Agencia Notimex por Plagio de 'Verificado'", 1 July 2019, https://vanguardia.com.mx/articulo/acusan-agencia-notimex-por-plagio-de-verificado; and Christina Tardáguila, "Lopez Obrador Launches its Own 'Verificado' and Infuriates Fact-Checkers in Mexico", 9 July 2019, Poynter, https://www.poynter.org/ifcn/2019/lopez-obrador-launches-its-own-verificado-and-infuriates-fact-checkers-in-mexico/.

[8] Government of India Press Information Bureau, PIB Fact Check, https://pib.gov.in/factcheck.aspx; The Indian Express, "PIB to Fact-Check Govt-Related News", 29 November 2019, https://indianexpress.com/article/india/pib-to-fact-check-govt-related-news-6141924/; and Smriti Kak Ramachandran, "PIB Plans a Fact-Checking Unit to Counter Fake News", 3 July 2019, Hindustan Times, https://www.hindustantimes.com/india-news/pib-plans-a-fact-checking-unit-to-counter-fake-news/story-BwNk8Y0TTj5WThE2Cy8BFI.html.

[9] @FakeNews_Buster, https://twitter.com/fakenews_buster?lang=en; and Dawn, "Govt Launches 'Fake News Buster' Account to Expose False Reports", 1 October 2018, https://www.dawn.com/news/1436167.

[10] The body is established by Ley 26.522, 10 October 2009, http://defensadelpublico.gob.ar/wp-content/uploads/2016/04/ley26522.pdf.

[11] See, for example, Directorio Legislativo, *NODIO: An Observatory Created by the Government to Control Media Misinformation*, 2020, https://directorio.cloud/csg-wp/wordpress/index.php/2020/10/26/no-dio-an-observatory-created-by-the-government-to-control-media-misinformation-newsletter/; and Clarín, "La Sociedad Interamericana de Prensa Repudió la Creación del Observatorio Nodio", 13 October 2020, https://www.clarin.com/politica/sociedad-interamericana-prensa-repudio-creacion-observatorio-nodio_0_5JZZnx7r4.html.

[12] Real 411, Media Monitoring Africa, https://www.real411.org/about; The Economist, "Censorious Governments

It may be preferable for official units focusing on disinformation to focus on disseminating accurate information to the public, especially if they are not independent of government. If so, it is important to impose clear mandates on these bodies. The United Kingdom, for example, created a Rapid Response Unit to address disinformation. The Unit is not independent, since it is housed within the Cabinet Office. Initially, the Unit stated that it was not an "anti-fake news" or fact checking body, instead serving as a Unit for tracking information trends and helping the government to develop appropriate messaging strategies to respond to inaccuracies.[13] However, news articles refer to the Unit as an "anti-fake news" unit, suggesting that the public perceives of it as playing this role.[14] In addition, during the pandemic, the Unit appears to be taking a more active role, such as working with social media companies on removing disinformation.[15] Such roles are better managed by a more independent body. At a minimum, a clearer official mandate could have prevented such drift.

Where official anti-disinformation or anti-misinformation units (i.e. with a mandate beyond disseminating accurate information) are created, they should have strong guarantees of their independence, clear mandates, appropriate powers and transparent rules regarding their mandates, powers and operating procedures. Strict rules should be in place to ensure that any power to correct inaccurate information conforms to the right to freedom of expression (which would allow for this only in very limited circumstances). Other powers, such as to tag information as incorrect or even issue counter-messaging, should be done by the public authority which has specialised expertise in the relevant area, such as public health authorities rather than a central authority for public health issues.

1.2 Legislative Responses

Laws prohibiting disinformation are not a new invention. France, for example, still has a prohibition on spreading false news in its Freedom of the Press Law, which dates from 1881.[16]

---

are Abusing 'Fake News' Laws", 13 February 2021, https://www.economist.com/international/2021/02/13/censorious-governments-are-abusing-fake-news-laws; and Simnikiwe Mzekandaba, "SA Deploys Hi-Tech to Fight Covid-19 Disinformation", 16 April 2020, https://www.itweb.co.za/content/nWJad7be1YlvbjO1.

[13] United States Law Library of Congress, "Government Responses to Disinformation on Social Media Platforms: United Kingdom", September 2019, https://www.loc.gov/law/help/social-media-disinformation/uk.php#_ftnref108.

[14] BBC, "Coronavirus: Fake News Crackdown by UK Government", 30 March 2020, https://www.bbc.com/news/technology-52086284; and Kate Proctor, "UK Anti-Fake News Unit Dealing with up to 10 False Coronavirus Articles a Day", 30 March 2020, The Guardian, https://www.theguardian.com/world/2020/mar/30/uk-anti-fake-news-unit-coronavirus.

[15] BBC, "Coronavirus: Fake News Crackdown by UK Government", *ibid*.

[16] Loi du 28 juillet 1881 sur la liberte de la presse, Article 27, https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006070722 (current version; original 1881 law can be downloaded from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000877119).

However, there had been somewhat of a trend towards repealing these older provisions as being incompatible with freedom of expression guarantees. For example, courts in both Zimbabwe and Uganda held that old colonial era laws prohibiting false news were unconstitutional, in 2000 and 2004 respectively,[17] while Zambia followed suit in 2014.[18] In 1992, Canada's Supreme Court found that a criminal prohibition on disseminating false statements likely to injure the public interest was unconstitutional because it violated the right to freedom of expression.[19]

Unfortunately, a surge of interest in disinformation and so-called "fake news", as well as the genuine challenges posed by the rapid spread of disinformation digitally, have more recently led to a series of new laws seeking to prohibit some variation of disinformation, misinformation, or false or misleading information. This is evidenced by a sampling of such laws passed in recent years:

- In April 2020, Algeria passed an amendment to its Penal Code which imposes up to three years' imprisonment and a fine for intentionally spreading false or slanderous information likely to harm public security or order. No definition of false information is provided.[20]
- Vietnam has had a prohibition on false news since it adopted its 2006 Law on Information Technology, but precise penalties for social media users were only clarified recently in Decree No. 15 of 2020, which imposes fines for sharing forged, false or distorted information over social networks. No further definition of what constitutes false or distorted information is given.[21]
- Nicaragua's Special Cybercrimes Law, which was enacted in October 2020, imposes between two and four years' imprisonment for publishing or disseminating false or distorted information online which produces alarm or fear. The penalty is up to three years' imprisonment for false information which damages a person's reputation, and up to five years' imprisonment where the information incites hate or endangers economic stability, public order, public health or sovereign security.[22]

---

[17] *Chavunduka v. Minister of Home Affairs*, 22 May 2000, Zimbabwe Supreme Court, SC 35/2000 (Zimbabwe Supreme Court), https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2017/08/Chavunduka-v-Minister-of-Home-Affairs-Zimbabwe9610.pdf; and *Charles Onyango Obbo and Another v. Attorney General*, 11 February 2004, Constitutional Appeal No. 2 of 2002 (Supreme Court of Uganda), https://ulii.org/ug/judgment/constitutional-court-uganda/2000/4.

[18] *Chipenzi and Others v. The People*, 4 December 2014, HPR/03/2014 (High Court for Zambia at Lusaka), https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2015/03/Chipenzi-v.-The-People-HPR032014.pdf.

[19] *R. v. Zundel*, 2 SCR 731, 27 August 1992 (Supreme Court of Canada), https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/904/index.do.

[20] Penal Code, Article 196-bis, as amended by Loi no. 20-06 modifiant et complétant l'ordonnance n° 66-156 du 8 juin 1966 portant code pénal, https://www.icnl.org/resources/library/loi-no-20-06-modifiant-et-completant-lordonnance-n-66-156-du-8-juin-1966-portant-code-penal.

[21] Decree No. 15/2020/ND-CP, 3 February 2020, Article 101(1)(a), translation on file with CLD.

[22] Ley N. 1042 (Ley Especial de Ciberdelitos), 27 October 2020, Article 30, http://legislacion.asamblea.gob.ni/normaweb.nsf/($All)/803E7C7FBCF44D7706258611007C6D87.

- Kenya's 2018 Computer Misuse and Cybercrimes Law included offences of "false publication" and "publication of false information". The former, which may result in a two year' imprisonment and/or a fine, prohibits publishing false or misleading data with the intent that it shall be acted upon as authentic. The latter addresses knowingly publishing information that is false and "is calculated or results in panic, chaos, or violence among citizens" or is likely to discredit the reputation of a person. Doing so may result in ten years' imprisonment and/or a fine.[23] The law itself is in legal limbo at the moment, however, due to a procedural issue surrounding its enactment.[24]
- Singapore's Protection from Online Falsehoods and Manipulation Act 2019 prohibits making false statements of facts which result in certain specified hams, including: prejudicing the security of Singapore; prejudicing public health, safety and tranquillity; influencing the outcome of an election; inciting feelings of enmity or ill-will between groups of persons; or diminishing public confidence in the government. The penalty for doing so is up to five years' imprisonment and/or a fine.[25]
- Ethiopia's 2020 Hate Speech and Disinformation Prevention and Suppression Proclamation criminalises disseminating disinformation, punishable by up to one year's imprisonment or a fine. This is increased to three years and the fine doubled if it is done through a social media account with more than 5,000 followers, a broadcaster or a print media outlet. Where violence or public disturbance results, the penalty is two to five years' imprisonment.[26] Disinformation is defined as speech that is false, where the person who is responsible knew or reasonably should have known it was false, and is "highly likely to cause a public disturbance, riot, violence or conflict".[27] Some exceptions apply for academic studies, news reports, political critique, artistic expression, religious teaching or where reasonable efforts where made to ensure accuracy.[28]
- In 2019, as part of an anti-terrorism agenda,[29] Burkina Faso amended its Penal Code to criminalise intentionally sharing false information suggesting that an attack or an act of

[23] Computer Misuse and Cybercrimes Law, No. 5 of 2018, sections 22-23, http://kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/ComputerMisuseandCybercrimesActNo5of2018.pdf.
[24] The false publication sections were already subject to a constitutional challenge which was on an appeal from a decision which declined to strike them down. However, the High Court also found that the entire law, along with a number of laws, had been enacted without proper involvement of the Senate. Article 19, "Court of Appeal's Ruling Strikes Further Blow to Free Expression and Privacy", 14 August 2020, https://www.article19.org/resources/kenya-court-of-appeals-ruling; and Parliament of Kenya, "National Assembly to Appeal High Court Ruling on Constitutional Petition", http://www.parliament.go.ke/national-assembly-appeal-high-court-ruling-constitutional-petition.
[25] Protection from Online Falsehoods and Manipulation Act 2019, No. 18 of 2019, section 7, https://sso.agc.gov.sg/Acts-Supp/18-2019.
[26] Proclamation No. 1185/2020, Articles 5, 7, https://www.article19.org/wp-content/uploads/2021/01/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf.
[27] *Ibid.*, Article 2(3).
[28] *Ibid.*, Article 6.
[29] Reuters, "Burkina Faso Urged to Avoid Curbs on Security Reporting", 3 July 2019, https://www.reuters.com/article/us-burkina-security-idUSKCN1TY2H6.

> property destruction has been or will be committed. The amendment defines false information as an inexact or misleading allegation or imputation of a fact.[30] The penalty is between one and five years' imprisonment and a fine.

Article 19(3) of the ICCPR defines a list of legitimate interests which restrictions on freedom of expression may protect, which is the rights and reputations of others, national security, public order, public health and public morals. Article 19(3) also requires restrictions to be "provided by law", which includes a requirement that they are clear and precise in nature, both so that individuals are warned in advance of what is prohibited and to prevent abuse of these provisions by officials. Finally, Article 19(3) also requires restrictions to be "necessary" which, among other things, means that they must be proportionate, in the sense of not harming freedom of expression to a greater extent than they deliver benefits through protecting the legitimate interest. This is a complex legal area, which has been interpreted to have different requirements in different areas, such as the system of defences to defamation claims and a requirement of a very close nexus between the expression and the risk of violence or harm to national security in those contexts.

Authoritative actors have held that it is not legitimate to impose general bans on the publication of incorrect or false statements. For example, the four special international mandates on freedom of expression[31] have adopted a Joint Declaration each year since 1999 and in 2017 it was the Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda. Paragraph 2(a) of that Joint Declaration stated:

> General prohibitions on the dissemination of information based on vague and ambiguous ideas, including "false news" or "non-objective information", are incompatible with international standards for restrictions on freedom of expression, as set out in paragraph 1(a), and should be abolished.[32]

All of the provisions above fail to respect one or another of these standards. To pass muster as a restriction on freedom of expression, a prohibition on false statements must, if criminal in nature, include an appropriate intent requirement. The absence of the latter means that the prohibition targets not just disinformation but also misinformation. This risks imposing severe penalties on those who are not even aware that they are sharing incorrect information or who have limited capacity to engage in fact checking, which is clearly disproportionate (and also fails to conform to standards relating to presumption of innocence).

[30] Loi N. 044-2019/AN, Articles 312-13, https://www.assembleenationale.bf/IMG/pdf/loi_044_portant_modification_du_code_penal.pdf.
[31] The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information.
[32] Adopted 3 March 2017, https://www.law-democracy.org/live/legal-work/standard-setting/.

Prohibitions on the dissemination of false statements linked to certain specific and clear harms, and where appropriate conditions and defences are provided, may be legitimate. Thus, every State bans false and defamatory statements, albeit international courts have made it clear that such laws are legitimate only if they make appropriate defences available. Similarly, prohibitions on intentionally lying in court (perjury) are ubiquitous.

The problem with the more general provisions noted above is that many fail to define clearly and sufficiently narrowly what resultant harm would trigger the offence (or, to put it another way, what legitimate interest is being protected), while some fail to tie the false statement to any harm at all. The provision in question in Zimbabwe in 2000 was conditioned upon the false statement being "likely to cause fear, alarm or despondency among the public" or "to disturb the public peace". The Supreme Court of Zimbabwe held that these were not sufficiently clear, did not link to a recognised interest (individuals do not have the right to be protected in general against fear and alarm) and, in addition, that there was not a sufficiently close link between the expression and the risk of harm.

Many of the provisions noted above include long lists of results that render the expression culpable, a sort of shopping list approach, with some items on the list appearing to be more legitimate than others. But even where the protected interest is legitimate – such as against violence or disorder – the rules fail to require a sufficient link between the statement and that result, again failing to strike an appropriate balance between freedom of expression and the competing interest.

The COVID-19 pandemic has significantly concentrated attention on the challenge of disinformation in the health context. Accordingly, during the COVID-19 pandemic, a number of States have introduced or amended additional rules around disinformation. Some examples include:

- In regulations under the Disaster Management Act, South Africa introduced criminal penalties of up to six months' imprisonment and/or a fine for publishing statements with the intention of deceiving another person about COVID-19, the infection status of a third party or any measure taken by the government to address COVID-19.[33]
- Thailand has had an Emergency Decree on Public Administration in Emergency Situation in place since 2005 which enables the Prime Minister to issue regulations which prohibit the publication of any communications "intended to distort information which misleads understanding of the emergency situation to the extent of affecting the security of state or public order or good moral of the people".[34] During the pandemic, Thailand has relied on this decree to prohibit reporting or spreading COVID-19 information which is untrue and may cause public fear. Officials are empowered to block such news or prosecute under

[33] Regulations Issued in Terms of section 27(2) of the Disaster Management Act, 2002, 18 March 2020, section 11(5), http://www.gpwonline.co.za/Gazettes/Gazettes/43107_18-3_COGTA.pdf.
[34] Emergency Decree on Public Administration in Emergency Situation, B.E. 2548 (2005), section 9(3), http://web.krisdika.go.th/data/document/ext810/810259_0001.pdf.

related provisions in the Computer-Related Crime Act or under the 2005 Emergency Decree.[35]

- In the Philippines, a March 2020 law granting the President powers to respond to the pandemic included a prohibition on false COVID-19 information. Specifically, the spreading of "false information regarding the COVID-19 crisis on social media and other platforms, such information having no valid or beneficial effect on the population, and are clearly geared to promote chaos, panic, anarchy, fear, or confusion" could be penalised by two months' imprisonment and/or a fine.[36] When this law expired, however, its replacement law did not include the prohibition on false information.[37] Prior to COVID-19, the Philippines' Penal Code already prohibited the publication of "false news which may endanger the public order, or cause damage to the interest or credit of the State".[38]
- In Romania, an emergency decree signed early in the pandemic allowed the communications regulatory authority to order website takedowns in response to COVID-19 disinformation. The OSCE Representative on Freedom of the Media voiced concerns over the fact that this provision allowed the takedown of entire websites and failed to provide for any appeal or redress mechanism, thereby unduly restricting the work of the media.[39]
- Bolivia, in two decrees responding to COVID-19, prohibited the dissemination of disinformation or information which generated uncertainty in the population, on the basis that this was a crime against public health.[40] A subsequent decree expanded this to *any* information which affects or puts at risk public health, dropping the "disinformation" requirement.[41]
- Botswana, in regulations enacted under the Emergency Powers Act in response to the pandemic, introduced a fine or up to five years imprisonment for anyone who relays *any*

[35] Human Rights Watch, "Thailand: State of Emergency Extension Unjustified", 27 May 2020, https://www.hrw.org/news/2020/05/27/thailand-state-emergency-extension-unjustified (linking to the list of prohibitions in Thai and summarising them in English).
[36] Republic Act No. 11469 ("Bayanihan to Heal as One Act"), 24 March 2020, section 6(f), https://www.officialgazette.gov.ph/downloads/2020/03mar/20200324-RA-11469-RRD.pdf.
[37] See Republic Act No. 11494 ("Bayanihan to Recover as One Act), 11 September 2020, https://www.officialgazette.gov.ph/downloads/2020/09sep/20200911-RA-11494-RRD.pdf.
[38] Revised Penal Code of the Philippines, Article 154(1), https://www.un.org/Depts/los/LEGISLATIONANDTREATIES/PDFFILES/PHL_revised_penal_code.pdf (fines were increased from this version by Republic Act No. 10951, 24 July 2017, section 18, https://www.lawphil.net/statutes/repacts/ra2017/ra_10951_2017.html).
[39] OSCE Representative on Freedom of the Media, "Coronavirus Response Bill Should Not Curb Freedom of Information in Romania, Stresses OSCE Media Freedom Representative", 30 March 2002, https://www.osce.org/representative-on-freedom-of-media/449380.
[40] Decreto Supremo N. 4200, 25 March 2020, Article 13(II), http://www.gacetaoficialdebolivia.gob.bo/edicions/view/1250NEC; and Decreto Supreme N. 4199, 21 March 2020, Article 7(II), http://www.gacetaoficialdebolivia.gob.bo/edicions/view/1249NEC.
[41] Decreto Supremo N. 4321, 7 May 2020, Article II, http://www.gacetaoficialdebolivia.gob.bo/edicions/view/1266NEC.

> information about COVID-19 to the public that is not from the government health services or the World Health Organisation (regardless of truth or falsity).[42]

The international law principles outlined above continue to apply during a health emergency. Specifically, even during an emergency, Article 19(3) continues to apply. States may, under certain limited conditions, derogate from certain human rights obligations, including the right to freedom of expression, during very serious emergencies. However, a study conducted by the Centre for Law and Democracy, focusing on the COVID-19 pandemic, concluded that there was fairly limited scope for this beyond the parameters of Article 19(3).[43] While it is important that people have access to reliable health information, it is also important that the public is able to criticise and disagree with government responses to the public, to dispute official statements, including as to the impact of the disease, and otherwise to hold the government to account for what are likely to be incredibly impactful decisions and actions that it is taking. Furthermore, given that misinformation rather than disinformation is often the most significant challenge in public health contexts, the efficacy of criminal penalties or fines, as opposed to educational campaigns and clear communication from public health authorities, is questionable.

Many of the COVID-19 responsive false news laws suffer from similar flaws to those identified above and some are even more problematical, essentially representing an attempt to give the government a monopoly over COVID-19 related information.

1.3 Requiring Action by Private Actors

Some governments have attempted to impose obligations on private sector actors, especially social media platforms, to take steps to combat disinformation. In milder forms, this can include softer non-binding negotiations or agreements with companies. However, some countries have attempted to impose legal liability on intermediaries for failing to act against disinformation disseminated through their services or present on their platforms. For example, the Ethiopian law on disinformation, referenced above, requires intermediaries to remove offending content within 24 hours of receiving notice that the content is illegal; the law is not clear on who issues this notice.[44]

A particularly well-known example is Germany's Network Enforcement Act, known by the German acronym NetzDG. NetzDG was introduced in the context of public concern in Germany

[42] Emergency Powers (COVID-19) Regulations, 2020, Statutory Instrument No. 61 of 2020, section 30(3), https://covidlawlab.org/wp-content/uploads/2020/06/Emergency-Powers-COVID-19-Regulations-2020.pdf.
[43] For a discussion of the legal issues around this, see the first part of Centre for Law and Democracy, *Maintaining Human Rights during Health Emergencies: Brief on Standards Regarding the Right to Information*, May 2020, https://www.law-democracy.org/live/wp-content/uploads/2020/05/RTI-and-COVID-19-Briefing.20-05-27.Final_.pdf.
[44] Proclamation No. 1185/2020, note 26, Article 8.

over disinformation and reporting on it often describes it as an anti-fake news measure,[45] but the law does not actually prohibit disinformation. The law requires social media networks with more than 2 million users in Germany to meet reporting obligations about handling complaints on their platforms and to have in place a transparent procedure for handling such complaints.[46] This procedure must include blocking or removing "manifestly" unlawful content within 24 hours of receiving a complaint and all unlawful content within seven days, with some exceptions.[47] Unlawful content is defined by reference to specific provisions of the Penal Code, including defamation and hate speech provisions. None of these penalise disinformation, although one includes a prohibition of "treasonous forgery", which is limited to contexts of deceiving foreign powers in a manner that creates a risk to Germany's security.[48]

Despite this, a number of countries have cited the NetzDG model in developing their own laws on disinformation. For example, the Malaysian Communications and Multimedia Minister directly referenced NetzDG when Malaysia adopted its Anti-Fake News Act in 2018 (since repealed but under consideration for re-introduction after two changes in government).[49] The 2018 law imposed criminal penalties on anyone who maliciously creates, publishes or disseminates fake news.[50] It also imposed fines on anyone who has a publication containing fake news in his or her "possession, custody or control" and failed to remove it, broadly covering all intermediaries as well as other people and entities, instead of targeting social media platforms like NetzDG.[51]

---

[45] See, for example, a BBC article stating that Germany will start enforcing a law requiring social media sites "to remove hate speech, fake news and illegal material". BBC, "Germany Starts Enforcing Hate Speech Law", 1 January 2018, https://www.bbc.com/news/technology-42510868.

[46] Act to Improve Enforcement of the Law in Social Networks (NetzDG), sections 1-3, English translation available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=AD99C47B2608D12B014859D5FF786F29.2_cid289?__blob=publicationFile&v=2.

[47] NetzDG, *ibid.*, section 3(2).

[48] German Criminal Code, 18 November 1998 as amended 19 June 2019, section 100a, English translation available at: https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html; a provision on data forgery which would create a counterfeit or falsified document is also incorporated (section 269).

[49] Jacob Mchangama and Joelle Fiss, *The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship* (2019, Copenhagen, Justitia), p. 8, http://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf (quoting the Malaysian Minister); Al-Jazeera, "Malaysia Parliament Scraps Law Criminalising Fake News", 10 October 2019, https://www.aljazeera.com/news/2019/10/10/malaysia-parliament-scraps-law-criminalising-fake-news; and The Straits Times, "Malaysia to Discuss Proposal to Revive Anti-Fake News Act in Parliament", 16 November 2020, https://www.straitstimes.com/asia/se-asia/malaysia-to-discuss-the-revival-of-anti-fake-news-act-at-parliament.

[50] Anti-Fake News Act 2018, Act 803, 11 April 2018, section 4, https://cyrilla.org/en/document/i675szk3h8m?page=6.

[51] Anti-Fake News Act 2018, note 50, section 6.

Similarly, ahead of Singapore's adoption of the Protection from Online Falsehoods and Manipulation Act 2019, a ministerial green paper presented to Parliament cited NetzDG.[52] In addition to prohibiting fake news, as described above, the Act represents one of the farthest reaching laws targeting false information online. It enables "any Minister" to require intermediaries to remove false statements or issue corrections if the Minister believes that doing so is in the public interest.[53] Such orders can also apply extraterritorially.[54]

The Singaporean Act also grants the government a number of other powers, such as requiring intermediaries to restrict the accounts of user who share false information or to block websites which repeatedly publish false information from receiving donations or advertising revenues.[55] It also enables the creation of "codes of practice" which can require designated intermediaries to take actions such as reporting any suspicion of "coordinated inauthentic behaviour" or undertaking due diligence as regards false information.[56] Codes of practice have already been issued under this provision on topics such as the Code of Practice on Giving Prominence to Credible Online Sources of Information and the Code of Practice for Transparency of Online Political Advertisements.[57]

NetzDG and similar laws engage broader intermediary liability issues. Laws which require intermediaries to remove content without proper procedural protections have long been a problem in many countries with speech-restrictive laws. As noted by the special international mandates on freedom of expression in their 2011 Joint Declaration on Freedom of Expression and the Internet, laws that require intermediaries to screen content for legality and impose liability absent a judicial or analogous order to takedown content present problems in terms of freedom of expression.[58] Any intermediary liability scheme needs to be limited to content that can properly be restricted according to the test laid out in Article 19(3) of the ICCPR, including in terms of clarity and precise about what is covered, and include due process protections for users.

The laws described in this section do not meet these requirements. Singapore's law, for example, clearly breaches freedom of expression guarantees in terms both of the content it covers and the powers it allocates to government. In terms of the latter, it gives extraordinary powers to Ministers to determine what constitutes false information and to impose wide-ranging measures to restrict access to that information or to sites or services which carry that information. NetzDG is more restrained, but it still creates a regime which requires intermediaries to remove content to avoid liability without any order from a judicial or other authoritative body confirming that the content

[52] Jacob Mchangama and Joelle Fiss, note 49, p. 9.
[53] Protection from Online Falsehoods and Manipulation Act 2019, Act No. 18 of 2020, section 20, https://sso.agc.gov.sg/Acts-Supp/18-2019.
[54] *Ibid.*, section 25.
[55] *Ibid.*, sections 36-38 and 40.
[56] *Ibid*, section 48.
[57] POFMA Office, Codes of Practice, https://www.pofmaoffice.gov.sg/regulations/codes-of-practice.
[58] 2011 Joint Declaration on Freedom of Expression and the Internet, 1 June 2011, para. 2(b), https://www.law-democracy.org/live/legal-work/standard-setting/; see also 2017 Joint Declaration, note 32, para. 1(d).

is illegal. A full exploration of intermediary liability is beyond the scope of this Submission but, in the context of disinformation, a few points are relevant:

- First, any obligations imposed on intermediaries which rely on underlying content restrictions which are themselves not human rights compliant cannot be legitimate and will magnify the scope of breaches to the right to freedom of expression.
- Second, while there is a need to find solutions to the spread of disinformation online, this needs to be done in a manner that respects international standards governing intermediary liability, which are themselves based on human rights, including freedom of expression. Criminal and civil law restrictions on content require proof of wrongdoing before any sanctions or other remedies may be imposed. In contrast, imposing liability on social media platforms for third party content which ultimately turns out to have been illegal, in advance of giving them authoritative notice (i.e. from a judicial or quasi-judicial body) that the content is illegal, signally lacks these protections, thereby putting freedom of expression at risk. In this context, intermediaries will inevitably be over-inclusive in their removal of content so as to insure themselves against a risk of future liability (i.e. they will remove anything they believe a judge could decide is illegal, rather than just what a judge actually has decided is illegal). This problem is dramatically compounded where legal rules incorporate unclearly terms such as "fake news" or "misleading information". Requiring intermediaries to be transparent in decision-making about content removal, to establish clear protocols for this and to respect basic due process rules are all positive steps, but they cannot remedy the core problem described above.
- Third, the charged language around "fake news" in public discourse has promoted overreach as well as confusion regasrding efforts to combat it. The frequent references to NetzDG as a "fake news" law, and the reliance of governments in other countries on this law to enact laws which bear little resemblance to Germany's law, highlight the need for clarity in this area.

The European Union is in the process of developing a Digital Services Act (DSA) which would address a range of regulatory issues impacting online intermediaries. The current proposal has a complex intermediary obligation regime which, on the one hand, requires "notice and action" schemes but, on the other, includes limitations on intermediary liability via "liability exemption" categories and a "Good Samaritan" clause which gives partial protection from liability for intermediaries who undertake actions to identify problematic content on their own.[59]

[59] European Commission, Proposal for the Regulation on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 15 December 2020, https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-european-parliament-and-council-single-market-digital-services-digital. Chapter II addresses liability issues while Chapter III, section 2, addresses notice and action mechanisms. For a discussion of the issues at stake, see Svea Windwehr and Christoph Schom, "Our EU Policy Principles: Procedural Justice", 27 July 2020, EFF, https://www.eff.org/deeplinks/2020/07/our-eu-policy-principles-procedural-justice; and Article 19, "At a Glance: Does the EU Digital Services Act Protect Freedom of Expression", 11 February 2021, https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/.

In addition, the Digital Services Act proposed by the European Union would introduce new transparency reporting requirements, including in relation to content monitoring and automated content moderation, which could enable researchers and others to understand better how disinformation trends on certain platforms.[60] It would also require enabling users to identify the source of advertisements on online platforms.[61]

The DSA does not define any specific types of problematic content, such as disinformation, since individual States are responsible for any specific content prohibitions. It will therefore be very important to track how different States intend to or eventually apply the DSA's requirements in the context of local disinformation rules. The experience of NetzDG and related laws suggests that governments which are anxious to take action on disinformation are likely to interpret the notice and action requirement, along with ambiguities in the intermediary liability rules, in a manner that fails to respect freedom of expression. On the other hand, the transparency obligations in the DSA could enable better tracking and identification of disinformation, along with the development of more effective strategies for combatting it.

1.4 Internet Shutdowns

An especially drastic response to disinformation is to impose limited restrictions on certain Internet or mobile phone services that enable Internet access or even to shutdown Internet access entirely. India is the most prominent example of this approach: it commonly uses shutdowns as a tactic for responding to social unrest or as a response to protests and often cites disinformation as a justification for these shutdowns.[62] However, other countries, such as Ethiopia, Nigeria and Sri Lanka, have also cited disinformation as necessitating Internet shutdowns, typically as part of an attempt to block the spread of information believed to incite communal violence.[63]

Ethnic or communal violence is a very serious concern and efforts are needed to address disinformation which incites such violence, which is often associated with widespread hate speech, an important topic beyond the scope of this Submission. However, Internet shutdowns are not an appropriate response to this problem for several reasons.

[60] Proposal for a Digital Services Act, note 59, Article 23.
[61] *Ibid.*, Article 24.
[62] Access Now, "Targeted, Cut Off, and Left in the Dark", #KeepItOn, 2019, https://www.accessnow.org/cms/assets/uploads/2020/02/KeepItOn-2019-report-1.pdf; and Jayshree Bajoria, "Internet Clampdown Will Not Stop Misinformation", Human Rights Watch, 24 April 2019, https://www.hrw.org/news/2019/04/24/india-internet-clampdown-will-not-stop-misinformation.
[63] Access Now, "The State of Internet Shutdowns Around the World", #KeepItOn, 2018, https://www.accessnow.org/cms/assets/uploads/2019/07/KeepItOn-2018-Report.pdf.

First, although governments often say shutdowns are designed to respond to disinformation, in reality this is often simply a convenient excuse and not the real motivation for these actions.[64] Sometimes, the reference to "disinformation" is specifically intended to distract focus from government conduct. An analysis of shutdowns in India, for example, found that the reasons for Internet shutdowns were more likely to be State violence (such as police misuse of force) than misinformation.[65] Second, shutdowns are questionably efficacious at reducing violence. There is limited research on their impact in this regard but at least one study in India found that shutdowns targeting protests were associated with *increased* collective violence as compared to protests where Internet access was maintained, perhaps because peaceful organising was challenging without Internet, speech countering the disinformation was also blocked or people did not have access to information about how to protect themselves.[66]

Importantly, from a human rights law perspective, Internet shutdowns are a wholly disproportionate response and cannot pass the ICCPR Article 19(3) test. By limiting access to all types of online information, including that which could potentially correct the disinformation in question, and given the widespread impact shutdowns have on basic service provision, including life-saving medical and humanitarian services, the cost of shutdowns will exceed any potential benefit. Ultimately, Internet shutdowns simply distract from finding solutions to the underlying social issues which provoke unrest, at a very steep cost which cannot be justified under international human rights law.[67]

1.5 Rights-Respecting Approaches: Recommendations

There is no quick fix response to disinformation. Instead, any effective response is will require substantial, long-term engagement. Ultimately, people are less susceptible to disinformation when they have access to and rely regularly upon trusted sources of information and when their capacity to identify and avoid disinformation is bolstered. Such sources should normally include official government sources, the media and others, such as civil society. Governments should make it a priority to preventing their own offices, officials, spokespersons and political representatives from

[64] Access Now, note 62, p. 39.
[65] Nehal Johri, "India's Internet Shutdowns Function Like 'Invisibility Cloaks'", Deutsche Welle, 13 November 2020, https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554.
[66] Jan Rydzak, "Of Blackouts and Bandhs: The Strategy and Structure of Disconnected Protest in India", 11 February 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330413 .
[67] Concerns with shutdowns along these lines have regularly been raised by the Special Rapporteur in his reports. See, for example, Special Rapporteur for Freedom of Expression, *Report on Freedom of Expression, States and the Private Sector in the Digital Age*, 11 May 2016, para. 48, https://www.undocs.org/A/HRC/32/38; Special Rapporteur for Freedom of Expression, *Report on the Role of Digital Access Providers*, 30 March 2017, paras. 8-16, https://www.undocs.org/A/HRC/35/22; and Special Rapporteur for Freedom of Expression, *Report on Contemporary Challenges to Freedom of Expression*, 6 September 2016, para. 21-22, https://www.undocs.org/A/71/373.

disseminating disinformation. They should also work to build trust among the public and enhance their provision of reliable, non-politically slanted information. A third vector in this regard is to improve the quality and implementation of right to information laws, which provide a citizen-led means for accessing government information. Governments should also take effective steps to promote media diversity and a strong flow of professional content to the public via the media. This should include measures to promote media sustainability, under serious strain today due to the double impact of the migration of audiences and advertising revenues online and the impact of the COVID-19 pandemic, and to support independent public service broadcasters, which can offer trustworthy content to the public for free.

Any laws which prohibit the dissemination of inaccurate information should be carefully tailored, as outlined above, so as to meet the standards of Article 19(3) of the ICCPR. Blanket, overbroad or unclear prohibitions on inaccurate statements are not appropriate. Laws which address disinformation in specific contexts, such as elections, public health or consumer safety, may be legitimate, as long as they are carefully drafted and provide an effective response to the problem. Many countries already have laws addressing issues such as truth in product advertising, for example, which responds to an important need. Elections is another area efforts have been made, especially in Western countries, to introduce rules to reduce the impact of disinformation on elections.[68] For the most part, however, anti-disinformation legislation tends to overbroad and used more as a tool to silence government critics than to protect human rights and other important social values.

The use of the criminal law to address disinformation should generally be re-evaluated. At a minimum, any such laws should, in addition to being narrowly drafted and linked to the prevention of specific harms, have strong and tailored intent requirements. These should go beyond simple intent to commit the act (i.e. disseminating disinformation) and include a more specific intent to create the harm sought to be avoided. Even then, care is needed. Ordinary people engaged in often rapid-fire debate on social media or other platforms, even if they are aware that information is inaccurate, may make poor decisions in the heat of the moment. Consideration should be given to limiting criminal penalties to cases where the means used to share the disinformation make it potentially more harmful, such as through the use of bots or anonymous accounts.

When it comes to requiring intermediaries to act against disinformation, great care should be taken not to impose legal obligations on them to monitor or takedown content before an authoritative and independence actor, such as a court or designated regulator, has declared it to be illegal. Content removal by intermediaries not only poses serious risks to freedom of expression but it also typically comes too late to actually be effective in terms of preventing the spread of disinformation.

[68] See, for example, Canada's Election Modernization Act, S.C. 2018, c. 31, 13 December 2018, https://canlii.ca/t/53jm9; and France's election law reforms targeting false information, including Loi n. 2018-1202, 22 December 2018, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037847559.

A positive approach is to require greater transparency on the part of intermediaries. This can apply not only to their own dedicated measures to address prohibited content, whether by law or their own terms of service, but also to features of their platforms, such as algorithms and artificial intelligence (AI) driven tools, that not only allow but actually promote the spread of disinformation. This is discussed further in the next section.

Greater attention should also be given to education and media and digital literacy efforts for both the general public and key actors who are well positioned to combat disinformation. While the precise impact of these efforts can be hard to measure, there is evidence that they contribute to limiting disinformation. Finland is often cited as having one of the highest rates of resistance to disinformation in Europe. This is partly due to a set of 2016 policies, which included introducing media literacy into the educational curriculum and piloted an inter-governmental committee which trained thousands of civil servants, teachers, librarians, welfare organisations, police and others to combat disinformation.[69] Training programmes also bear a low risk of being rights restrictive, provided they are run independently and are not co-opted for propaganda purposes.

Similarly, governments should carefully distinguish between disinformation and misinformation, and approach the latter as a public service delivery rather than a criminal issue. For example, Internews has developed guides on addressing rumours in humanitarian contexts that provide staff dealing with vulnerable groups with guidance on combating misinformation.[70] In the context of COVID-19, the World Health Organization has provided similar guidance to public health professionals.[71] Governments should consider how to equip actors working at the grassroots level, such as public health professionals and community service workers, with the tools to respond to dangerous rumours.

Recommendations:

- Any official bodies which have the power to go beyond mere monitoring of and reporting on disinformation should be strictly independent of government and have their mandate and powers clearly defined. At the same time, consideration should be given to providing support for independent anti-disinformation initiatives. Effective measures should be put in place to limit the extent to which any anti-disinformation measures are politicised.
- Governments should prioritise limiting the spread of disinformation by official actors.

[69] Jon Henley, "Finland Starts its Fight against Fake News in Primary Schools", 29 January 2020, The Guardian, https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news; and Reid Standish, "Why is Finland Able to Fend Off Putin's Information War?", 1 March 2017, Foreign Policy, https://foreignpolicy.com/2017/03/01/why-is-finland-able-to-fend-off-putins-information-war.
[70] Internews, *Managing Misinformation in a Humanitarian Context* (2019), https://internews.org/sites/default/files/2019-07/Rumor_Tracking_Mods_3_How-to-Guide.pdf.
[71] See, for example, World Health Organization, "Infodemic Management", https://www.who.int/teams/risk-communication/infodemic-management.

- They should also focus on measures to enhance the flow of accurate, reliable information to the public, including through:
  - Building trust with citizens and enhancing the flow of reliable, accurate information from official sources to citizens.
  - Ensuring the adoption and implementation of strong right to information laws.
  - Promoting a strong, independent and diverse media sector, including through supporting media sustainability initiatives and independent public service broadcasters.
- Any laws which penalise the dissemination of inaccurate information should be carefully and narrowly tailored to protect specific, legitimate interests against harm, including in specific contexts, such as elections or public health, in strict compliance with Article 19(3) of the ICCPR.
- Any criminal prohibitions on the dissemination of inaccurate information should include appropriate intent requirements.
- Consideration should also be given to limiting the application of any criminal prohibitions to those who use certain more high-powered tactics designed to create harm, such as the use of bots, rather than individual social media users.
- Internet shutdowns should never be used as a response to disinformation.
- Governments should promote educational and training efforts, including media and information literacy programmes for both the general public and other key actors who may be in a better position to limit the negative impact of disinformation. This should include a focus on supporting those working at the grassroots level to combat disinformation in the communities they work in.

2. Private Sector Responses to Disinformation

There have been a few joint initiatives to develop standards around disinformation in the private sector, particularly in relation to the large intermediaries which facilitate much of our modern communications. For example, in one of the more targeted, a State-led initiative, the European Commission has developed a Code of Practice on Disinformation to which companies can voluntarily commit and many large ones – including Facebook, Google and Twitter – have signed on. The Code addresses issues such the placement of advertisements, political advertising and integrity of services (such as monitoring for fake accounts). The signatories commit to not deleting content merely because it is false, acknowledge the importance of limiting the visibility of disinformation and make some transparency commitments around policies and decision-making.[72]

[72] European Commission, Code of Practice on Disinformation, 2018, https://ec.europa.eu/digital-single-market/en/code-practice-disinformation.

Generally, however, actions against disinformation by companies themselves operate at the individual company level, and depend on the particular company's choices and approach. This section briefly surveys disinformation policies and actions taken by the largest intermediaries, organised by company group. It then summarises some of the main ongoing challenges to combating disinformation despite these efforts and offers recommendations for reform.

2.1 Facebook/Instagram and WhatsApp

Facebook has a set of Community Standards which represent its own standards for behaviour on its services, including in relation to content. A version of these is public[73] but there is reportedly a more detailed version, used by internal content reviewers, which is not public.[74] The public version includes a section addressing "false news". This concept is not defined but the policy states that false news will not be removed because of the "fine line between false news and satire or opinion". Instead, Facebook will de-prioritise such news on its News Feed.[75]

Facebook's most notable anti-disinformation initiative is its fact-checking programme, which is operated by independent third-party fact checkers. These fact checkers are responsible for identifying misinformation. Once flagged, Facebook says it will then label the content, de-prioritise it on the news feed and other features and/or highlight "related articles" which direct the user to trustworthy articles. Facebook lists the fact checkers it relies on in each country, which are certified by the International Fact-Checking Network,[76] and provides information on how publishers can contest ratings by emailing the third-party fact checker who made the decision.[77]

Facebook does also prohibit and may remove other content which has a bearing on disinformation. "Inauthentic behaviour", meaning misleading persons as to identity or purpose of an account or its origins or sources, is prohibited, as is "manipulated media" such as videos doctored so as to be misleading.[78] The Standards also prohibit "coordinating harm", which includes promoting "harmful miracle cures" for health issues and misrepresentation of certain information about

[73] Facebook, Community Standards, https://www.facebook.com/communitystandards/.
[74] Andrew Marantz, "Why Facebook Can't Fix Itself", 12 October 2020, The New Yorker, https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself.
[75] Facebook, Community Standard 21: False News, 2021, https://www.facebook.com/communitystandards/false_news.
[76] Poynter, The International Fact-Checking Network, https://www.poynter.org/ifcn.
[77] Facebook for Business, Issue a Correction or Dispute a Rating, 25 February 2021, https://www.facebook.com/business/help/997484867366026?id=673052479947730; and Facebook for Business, Fact-Checking on Facebook, 14 December 2020, https://www.facebook.com/business/help/2593586717571940?id=673052479947730&recommended_by=997484867366026.
[78] Facebook, Community Standard 20: Inauthentic Behavior, 2021, https://www.facebook.com/communitystandards/inauthentic_behavior; and Facebook Community Standard 22: Manipulated Content, 2021, https://www.facebook.com/communitystandards/manipulated_media.

voting and the census which may constitute voter and/or census interference, such as providing misleading information about where to vote.[79]

The public Community Standards are mostly very general in nature. However, in the context of COVID-19, Facebook has introduced much more comprehensive guidance on what content it will remove. Facebook's COVID-19 and Vaccine Policy states that it will remove misinformation if the information is false, according to public health authorities, and likely to contribute to imminent physical harm. The Policy contains a detailed list of what claims about COVID-19 may be removed, along with specific examples of statements that are not permitted.[80]

Facebook also reports that it is taking additional steps to combat disinformation, such as removing financial incentives for types of content which often feature misinformation, such as clickbait or low quality web pages primarily designed to generate ad revenue.[81] Ahead of 2020 elections in the United States, it announced steps such as placing clearer warning labels on posts identified as false.[82] In 2020, it also reportedly experimented with a "circuit breaker" approach which limits the spread of potentially harmful viral stories early on, before they spread, while the company decides whether they violate its policies.[83]

In addition to the third-party fact checkers, Facebook also has an internal content review team which is generally responsible for enforcing Community Standards.[84] Facebook has not always been clear about how the dual review systems work and interact, especially in controversial contexts. For example, Facebook's policy states that independent fact checkers do not fact check statements by politicians, given the importance of such statements to open democratic debate.[85] However, since 2020, in a well publicised policy shift, Facebook announced that it would add warning labels to posts by politicians, in response to controversy over its decisions not to flag posts by United States President Donald Trump.[86] Despite this change, there has been no change in the

[79] Facebook, Community Standard 3: Coordinating Harm and Publicizing Crime, 2021, https://www.facebook.com/communitystandards/coordinating_harm_publicizing_crime.
[80] Facebook, COVID-19 Policy and Vaccine Policy Updates and Protections, 2021, https://www.facebook.com/help/230764881494641.
[81] Tessa Lyons, "Hard Questions: What's Facebook's Strategy for Stopping False News?", Facebook, 23 May 2018, https://about.fb.com/news/2018/05/hard-questions-false-news.
[82] Guy Rosen. *et al.*, "Helping to Protect the 2020 US Elections", 21 October 2018, Facebook, https://about.fb.com/news/2019/10/update-on-election-integrity-efforts.
[83] Jeff John Roberts, "Facebook's New Tool to Stop Fake News is a Game Changer- if the Company Would Only Use It", 18 October 2020, Fortune, https://fortune.com/2020/10/18/facebook-tool-stop-fake-news-viral-content-review-system-fb-business-model.
[84] Facebook, Understanding the Community Standards Enforcement Report, 2021, https://transparency.facebook.com/community-standards-enforcement/guide#section3.
[85] Facebook for Business, Program Policies, 12 August 2020, https://www.facebook.com/business/help/315131736305613?id=673052479947730.
[86] Shannon Bond, "In Reversal, Facebook to Label Politicians' Harmful Posts as Ad Boycott Grows", NPR, 26 June 2020, https://www.npr.org/2020/06/26/883941796/unilever-maker-of-dove-soap-is-latest-brand-to-boycott-

relevant part of the publicly posted policy on fact checkers, which suggests that reviews of posts by politicians must be done by the internal review team rather than independent fact checkers, although this is not clear. Part of the problem here is that while Facebook produces transparency reports on content actions, these do not include information on the Community Standards categories which are relevant to disinformation.[87]

Overall, Facebook's approach to disinformation relies on different types of content moderation. In some cases, content is "controlled", either by being removed or de-prioritised (rendered less visible), while in other cases a "more speech" approach, such as labelling and directing to other content, is employed. Because the control measures place Facebook in the position of deciding what speech is acceptable, they raise serious concerns from a freedom of expression standpoint. This is exacerbated by the very general language of Facebook's Community Standards. Facebook has shown, with its standards on COVID-19 misinformation, that it can generate far more precise standards on what speech it will and will not moderate, which in turn enables more sophisticated public debate about whether those standards are appropriate. To develop more precise standards for its Community Standards in general, which we recommend, Facebook should draw on international human rights law. This can provide guidance on how to balance freedom of expression and other social interests in "hard cases", which the current Community Standards are simply too ambiguous to do properly.

The free speech implications of Facebook's work on disinformation is also exacerbated by the lack of transparency around internal decision-making. There is no reason for Facebook not to make its more detailed internal rules public and not doing so raises serious questions about Facebook's commitment to transparency in the area of content moderation.

The efficacy of Facebook's content moderation and fact checking is also questionable, because it typically comes too late, after disinformation has already spread. Far more effort needs to be focused on the extent to which Facebook's general business model facilitates or even promotes the sharing of disinformation in the first place. Facebook has shown some engagement with other solutions which may have a greater impact on its business than the content moderation approaches do, such as seeking to limit the spread of misinformation through content designed primarily to generate ad clicks. Far more serious engagement at that level is needed.

Facebook also owns messaging service WhatsApp. WhatsApp communications have end-to-end encryption and it is therefore not possible for it to monitor content directly. However, WhatsApp has taken some measures to limit behaviour deemed to be linked to disinformation. In particular, it has progressively taken steps to limit the number of times messages can be forwarded. As of April 2020, once a message has been forwarded frequently (defined as having been forwarded five

facebook.

[87] Facebook, Community Standards Enforcement Report, February 2021, https://transparency.facebook.com/community-standards-enforcement.

or more times) can only be further forwarded to a single chat at a time and will have a "frequently forwarded" icon attached to them. This followed a 2019 restriction dropping the number of permissible forwards for any message from twenty to five.[88] WhatsApp has indicated that, in the weeks after the 2020 change, they saw a 70% drop in frequently circulated messages, although there is no way to know what proportion of those contained false statements.[89]

This approach allows WhatsApp to take some steps to address disinformation without breaching its end-to-end encryption, which is a key feature for protecting freedom of expression and privacy, including for human rights defenders who fear surveillance on account of their work. However, because it applies randomly to all messages, the restriction captures non-harmful and even public interest speech as well as disinformation. It seems unlikely that guarantees of freedom of expression would allow a State to take such an untargeted approach, although this may depend on a weighing of the benefits as compared to the harm, as well as an assessment of what other options might be possible. WhatsApp has a responsibility to exercise due diligence in exploring other technically possible options for more targeted solutions to address disinformation while preserving end-to-end encryption. For example, a WhatsApp-funded report on the role of WhatsApp in spreading misinformation in India recommended a "beacon" feature, which would broadcast warnings or advisories to users in specific locations, a "more speech" as opposed to a control of speech approach.[90]

2.2 Twitter

Twitter's rules include a "civic integrity policy" which prohibits the use of Twitter for manipulating or interfering in elections or other civic processes. Twitter will label or remove false or misleading information about how to participate in elections or civic processes, information that is misleading and intended to suppress participation in such processes or misleading information about their outcomes. However, statements that are merely inaccurate, polarising or partisan will not violate the policy.[91]

[88] Alex Hern, "WhatsApp to Impose New Limit on Forwarding to Fight Fake News", 7 April 2020, The Guardian, https://www.theguardian.com/technology/2020/apr/07/whatsapp-to-impose-new-limit-on-forwarding-to-fight-fake-news.
[89] Manish Singh, "WhatsApp's New Limit Cuts Virality of 'Highly Forwarded' Messaged by 70%", Tech Crunch, 27 April 2020, https://techcrunch.com/2020/04/27/whatsapps-new-limit-cuts-virality-of-highly-forwarded-messages-by-70.
[90] Shakuntala Banaji and Ram Bhat, WhatsApp Vigilantes: An Exploration of Citizen Reception and Circulation of WhatsApp Misinformation Linked to Mob Violence in India, 2019, https://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf.
[91] Twitter Help Center, Civic Integrity Policy, January 2021, https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

Other relevant policies include the "synthetic and manipulated media policy" which evaluates content based on three factors: whether the content is significantly and deceptively altered or fabricated, whether it is shared in a deceptive manner, and whether it is likely to cause serious harm to public safety.[92] Misleading other Twitter users via fake accounts is also prohibited under Twitter's platform manipulation and spam policy, with exceptions for things such as parody and fan accounts[93]

Twitter has also introduced a "Covid-19 Misleading Information Policy". Under this policy, Twitter may label or remove claims of fact, expressed in definitive terms, which are demonstrably false or misleading based on widely available, authoritative sources and which are likely to "impact public safety or cause serious harm". The Policy goes on to identity categories of content it will remove, such as false claims about the nature of the virus or about the efficacy of preventative measures and treatments. The Policy notes that strong commentary or opinions, satire, counter speech (i.e. responding to misleading claims), personal anecdotes and public debate about COVID-19 research will generally be protected.[94]

Where content violates Twitter's rules, it may take a range of actions, including limiting the visibility of a tweet, requiring the tweet's removal before the user is allowed to tweet again, hiding a tweet, placing an account in a temporary "read-only" mode where previous tweets remain visible but the user cannot issue new tweets or retweet, verifying account ownership or permanently suspending an account. In rare cases, Twitter may decide that there is public interest in a tweet that would normally violate the rules, in which case Twitter will add a notice to the tweet and limit engagement with the tweet, such as replies, retweets, and likes, but still allow users to view it.[95]

Twitter is also experimenting with platform features designed to slow disinformation. Twitter now encourages users to add a comment when retweeting content, in the hopes that they will actually read the content they are retweeting. It is also reportedly trialling a feature which would warn a user who likes a tweet labelled as misinformation, aiming to diminish the number of likes such content receives.[96] Another new initiative is "Birdwatch", introduced as a pilot in January 2021,

[92] Twitter Help Center, Synthetic and Manipulated Media Policy, 2021, https://help.twitter.com/en/rules-and-policies/manipulated-media.
[93] Twitter Help Center, Platform Manipulation and Spam Policy, September 2020, https://help.twitter.com/en/rules-and-policies/platform-manipulation.
[94] Twitter Help Center, COVID-19 Misleading Information Policy, 2021, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.
[95] Twitter Help Center, Range of Enforcement Options, 2021, https://help.twitter.com/en/rules-and-policies/enforcement-options.
[96] Sarah Perez, "Twitter May Slow Down Users' Ability to 'Like' Tweets Containing Misinformation", 9 November 2020, Tech Crunch, https://techcrunch.com/2020/11/09/twitter-may-slow-down-users-ability-to-like-tweets-containing-misinformation.

which enables crowd-sourced notes to provide additional context for tweets which contain misleading information.[97]

Twitter's approach to disinformation ultimately shares some of the same weaknesses as that of Facebook. While it has policies in place prohibiting disinformation, for the most part these policies are drafted in very general language, allowing for subjective interpretation, and appear to be driven by public pressure in relation to sensitive topics, which creates obvious risks for freedom of expression. The range of measures available to Twitter is also expansive and includes very speech intrusive measures such as taking tweets down and even banning accounts. A positive feature of Twitter's approach is its "public interest exception", which acknowledges the importance of keeping some information on its platform even if it violates Twitter's content rules. However, Twitter's rules on disinformation should be articulated more clearly and objectively, again with human rights law standards providing guidance. In addition, a key underlying problem is again a lack of clarity about how Twitter prioritises certain content and how this may contribute to the spread of disinformation. While technical experiments like "Birdwatch" are intriguing, they may not be enough to overcome structural features of Twitter which enable the spread of disinformation in the first place, and about which there is insufficient transparency.

2.3 Google and YouTube

Google Search only removes content from their search results rarely and does not have a policy of removing disinformation from search results, on the theory that all information should be available to users.[98] However, Google does state that its search tools prioritise "high-quality" content. Google works with external "Search Quality Raters" who evaluate search results according to a set of General Guidelines.[99] In 2016, guidance was added to these Guidelines to advise the Raters to give lower quality ratings to pages with inaccurate content.[100] The Guidelines now instruct Raters to classify pages in the "lowest category" if they have "demonstrably inaccurate content", "debunked or unsubstantiated conspiracy theories" or content that contradicts expert consensus on information which could harm a person's "future happiness, health, financial stability, or safety" ("Your Money or Your Life").[101]

[97] Keith Coleman, "Introducing Birdwatch, A Community-Based Approach to Misinformation", Twitter Blog, January 2021, https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.
[98] Google, "How Google Fights Disinformation", February 2019, p. 12, https://blog.google/documents/37/How_Google_Fights_Disinformation.pdf.
[99] Google, General Guidelines, 14 October 2020, https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf.
[100] Google, "How Google Fights Disinformation", note 98, p. 13.
[101] Google, General Guidelines, 14 October 2020, note 99, pp. 10 and 44.

While Google typically will not remove content from search results altogether, it may remove or limit it in the context of specific features where the feature itself promotes certain content. One such example is the autocomplete feature which prompts users with search options as they type.[102] However, Google's Autocomplete Policies do not list disinformation or misinformation among the grounds for removing autocomplete suggestions.[103] Another more relevant example for disinformation is, therefore, Google's "featured snippets" policy, which places the short description of a search result before rather than after the page link. Google may remove featured snippets which do not respect its policies, such as those "contradicting consensus on public interest topics" including civic, medical, scientific and historical issues.[104]

Google News, unlike Google Search, will not feature content that violates its policies on its "news surfaces", while repeated breaches may lead to entire sites being removed. The policies cover cases of misrepresentation of the identity of the source of information or of inauthentic or coordinated behaviour that misleads users. Content which has been manipulated so as to mislead ("manipulated media") is also not allowed, along with medical content that contradicts scientific or medical consensus and evidence-based best practices.[105]

Google-owned YouTube will also remove content where it is in violation of its Community Guidelines; after three strikes a channel will be terminated. The Spam, Deceptive Practices and Scams Policies prohibit manipulated media that "may pose a serious risk of egregious harm", misleading thumbnails or metadata which lead users to believe content is about something different than it actually is.[106] The Guidelines also cover a number of types of misleading information related to voting or participating in the census or about the eligibility of candidates for office.[107] The Guidelines are applied by YouTube's own Community Guideline reviewers, who are responsible for removing videos that violate its rules.[108]

In addition, YouTube reports that it prioritises, via machine learning systems, information from authoritative sources in search results and in the recommended videos feature. It also states that it has "information panels" that provide additional information from "authoritative third-party sources" alongside videos containing types of content that often are accompanied by

[102] Google, "How Google Fights Disinformation", note 98, p. 12.
[103] Google Search Help, Autocomplete Policies, 2021, https://support.google.com/websearch/answer/7368877?p=blog_autocomplete_policies&visit_id=1-636287257258669056-1963168283&rd=1.
[104] Google Search Help, How Google's Featured Snippets Work, 2021, https://support.google.com/websearch/answer/9351707?hl=en#zippy=%2Cwhy-featured-snippets-may-be-removed.
[105] Google Publisher Center Help, Google News Content Policies, 2021, https://support.google.com/news/publisher-center/answer/6204050?hl=en.
[106] YouTube Help, Spam, Deceptive Practices, & Scams Policies, 2021, https://support.google.com/youtube/answer/2801973?hl=en.
[107] YouTube Help, Spam, Deceptive Practices, & Scams Policies, note 106.
[108] YouTube Help, Appeal Community Guidelines Actions, 2021, https://support.google.com/youtube/answer/185111?hl=en&ref_topic=9387060.

misinformation.[109] In making these decisions about what constitutes misinformation, YouTube uses external evaluators and experts. These experts, who provide feedback on search results, recommendations and relevance of videos, are trained using the same guidelines that Google uses to identity "lower quality" content in search results, described above.[110] YouTube then feeds the input from these evaluators into automated systems that review videos to identify misinformation.[111]

While Google Search and YouTube are quite different in terms of their focus and content, both have an enormous influence on what content users encounter through the content they prioritise: Google search by what search results list first and YouTube via the both the search and the "recommended videos" feature. While the General Guidelines provide relatively detailed instructions to Raters on how to go about the actual rating, very little guidance is given on how to decide what constitutes inaccurate content or debunked conspiracy theories, or when such content might still be of public importance versus when it may pose harms. Linking these definitions to almost impossibly vague notions such as "future happiness" is one aspect of this. Once again, clearer rules, guided by human rights law, should be introduced. In addition, while Google states that it uses external reviewers, it is not clear how independent the process is in practice and there is little transparency about the search tools used to prioritise content on Google and YouTube.

2.4 Summary of Ongoing Challenges and Recommendations

The main intermediaries have policies on content moderation, which include measures to limit the spread and impact of disinformation, and many have taken other steps, such as technical measures, to respond to disinformation. While this may seem to be quite engaging on the surface, a closer look reveals a number of ongoing challenges.

First, the public versions of their policies vary in terms of the level of detail and guidance they provide as regards what content is subject to moderation on grounds that it is misleading or inaccurate, but all are sufficiently generic to allow significant discretion and subjectivity in their application. This is very problematical at different levels, including freedom of expression, and it does not need to be the case. The response of these companies to COVID-19 has been to adopt far more detailed and precise guidance and rules on COVID-19 and public health misinformation, demonstrating that this can be done. Intermediaries should adopt far clearer rules on what

[109] YouTube, How Does YouTube Combat Misinformation?, Raising Quality Info, 2021, https://www.youtube.com/intl/ALL_ca/howyoutubeworks/our-commitments/fighting-misinformation/#raising-quality-info.
[110] YouTube, How Does YouTube Combat Misinformation?, *ibid*; YouTube Help, External Evaluators and Recommendations, 2021, https://support.google.com/youtube/answer/9230586.
[111] YouTube, How Does YouTube Combat Misinformation?, Determining Misinfo, 2021, https://www.youtube.com/intl/ALL_ca/howyoutubeworks/our-commitments/fighting-misinformation/#determining-misinfo.

constitutes misleading or inaccurate content (and other areas of content regulation). In doing so, they could benefit from international human rights standards, which have several decades of experience wrestling with hard questions when it comes to freedom of expression.

Second, companies are still not sufficiently transparent about how they address disinformation (and other content moderation systems). This applies to the definition of what constitutes policy-breaching content as well as their procedures and decision-making regarding how these content standards are applied. In most cases, internal procedures for making decisions on content moderation are not disclosed at all. In some cases, such as with Facebook, even the rules on how problematical content is defined appear to be subject to more detailed internal guidance, which is not made public. It is not clear what harm might arise from making such information public, perhaps apart from criticism regarding the quality of the standards and their manner of application, which should be encouraged, not avoided.

Third, there is still very little transparency around how automated processing of content (algorithms or programmes) is used on these platforms to prioritise and showcase content to users. This is a major challenge in terms of addressing disinformation. For many social media platforms, for example, a key design feature of these systems is to showcase content which engages the user, thereby inevitably prioritising controversial, and potentially inaccurate, content. If so, when companies claim they are taking measures to address disinformation on their services, this may just represent a minor vanguard action against the natural tendencies of their dominant automated processes. Secrecy is a huge problem for researchers and advocates here, since they have very limited access to data that would enable them to analyse the extent to which these automated systems do in fact promote disinformation and propose solutions.[112] Intermediaries view their automated systems as propriety information – and, indeed, they are often key to their competitive success – prompting them to be very resistant to transparency in this area. However, to combat disinformation seriously, intermediaries may need to re-think the underlying content prioritisation approaches of their core businesses and also consider how to allow independent actors, perhaps on a limited, confidential basis, to engage with their automated systems so as to provide independent input into these issues.[113]

[112] Irene V. Pasquetto, *et al.*, "Tackling misinformation: What researchers could do with social media data", 9 December 2020, Harvard Kennedy School, HKS Misinformation Review, https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/.

[113] As stated by a recent report on COVID-19 disinformation: "Social media platforms should go further in addressing coronavirus misinformation and disinformation by structurally altering how their websites function. For the sake of public health, social media platforms must change their product features designed to incentivize maximum engagement and amplify the most engaging posts over all others... Platforms must pair these changes with unprecedented transparency in order to enable independent researchers and civil society groups to appropriately study their effects." Erin Simpson and Adam Conner, "Fighting Coronovirus Misinformation and Disinformation: Product Recommendations for Social Media", 18 August 2020, Center for American Progress, https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation. See also Svea Windwehr and Christoph Schom, note 59 (on the need for more

Fourth, most of the large intermediaries which are most relevant for this issue are based in highly developed countries, with many having their headquarters in the United States. This tends to be reflected in their policymaking on how to address problematical content. This is evidenced by the strong focus of the policies described above on election-related interference. YouTube even has a specific line in its Community Guidelines related to content falsely claiming impropriety in the outcome of the 2020 United States' presidential election, although it has no general policy setting out how it will address content related to fraudulent election claims in other countries or contexts.[114] This should also include a greater effort to understand local factors driving disinformation, which is often linked to very localised political, social and economic phenomena. Intermediaries are not responsible for the underlying problems which spur on disinformation but they should allocate far more resources to engage meaningfully around them. Options here might include collaborating and consulting more closely with civil society in countries where disinformation through their services is more prevalent and designing public accountability and feedback mechanisms which are more responsive to local contexts.

It is also questionable whether companies are investing sufficiently in addressing disinformation in different markets. A 2018 BBC investigation into ethnic violence in Nigeria linked to fabricated incendiary posts circulated on Facebook found that, at the time, Facebook had only four full-time fact checkers to cover its 24 million users in the country.[115]

Fifth, this whole issue raises serious concerns from the perspective of freedom of expression and other human rights about the responsibilities of private companies, as well as the human rights obligations of States when engaging with (or putting pressure on) them. Ultimately, the measures put in place by intermediaries have an increasingly important impact on what speech is and is not allowed. This raises serious human rights accountability concerns about these private companies, and especially the few very dominant ones. Transparency has to be a starting point in this regard, but there are also issues regarding procedure and, ultimately, substance.

Sixth, more attention needs to be given to the range of measures that are available, including investing in brainstorming and developing potential new measures. Where "more speech" approaches, such as labelling and directing users to verifiable content, are effective, they should be prioritised over more speech controlling measures such as content de-prioritisation or removal, let alone banning users or entire sites, which would properly be labelled as prior censorship if it were undertaken by a State actor. This should include a greater focus on alternative solutions to content measures as a solution to disinformation, particularly around the design of platforms

information on automated decision making).

[114] YouTube Help, Spam, Deceptive Practices, & Scams Policies, 2021, https://support.google.com/youtube/answer/2801973.

[115] Yemisi Adegoke and BBC Africa Eye, "Like. Share. Kill.: Nigerian Police Say False Information on Facebook is Killing People", 13 November 2018, BBC, https://www.bbc.co.uk/news/resources/idt-sh/nigeria_fake_news.

themselves and features which facilitate the spread of disinformation. This may include, as noted above, a re-evaluation of existing automated features and increased efforts to identity external automated actions, such as the use of bots and other tactics, to spread disinformation.

Recommendations:

- Intermediaries should develop clear and precise standards to identify what content may be subject to measures, including to address disinformation, and avoid rules that are unclear or vague. Consideration should be given, in this regard, to taking guidance from international human rights standards.
- Intermediaries should be far more transparent about both their substantive and procedural rules and systems (i.e. how policies are applied and enforced) to address the promotion of disinformation through their services, which should cover both internal and independent structures for this.
- Intermediaries should take more muscular and effective steps to address the issue of a lack of transparency regarding their automated systems and the risk that the core design of these systems leads to promoting or incentivising, rather than impeding, disinformation. Where it exists, this problem needs to be acknowledged and serious steps taken to address. The lack of access to these automated solutions means that it is not possible for external actors to propose concrete ideas here but one option could include allowing a limited number of independent researchers to study company automated systems, on a confidential basis, and come up with possible solutions.
- Intermediaries should take steps to address any geographic and other biases in the way they address disinformation globally. This could include allocating adequate resources to this issue globally, based on where the problem is more serious, tailoring policies so as to respond to issues globally and not just the way problems manifest themselves in the United States and other highly developed countries, increasing engagement with their users and civil society groups in countries where disinformation is more ubiquitous and problematical, and appropriate tailoring of responses to the underlying motivations for disseminating disinformation and misinformation in different countries.
- Intermediaries should be sensitive to the freedom of expression and other human rights impacts of both disinformation and their responses to it. Overall, they should accept more responsibility for human rights abuses which are facilitated by their actions and explore ways of increasing their accountability for this. As part of this, intermediaries should focus more attention on identifying, applying and then assessing less speech controlling approaches to addressing disinformation, for example in favour of more speech approaches, as well as technical responses to automated means of disseminating disinformation.