Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

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

September 2020

BROADBAND COMMISSION
FOR SUSTAINABLE DEVELOPMENT

Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

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

Editors:

Kalina Bontcheva & Julie Posetti

Contributing authors:

Kalina Bontcheva	University of Sheffield, UK
Julie Posetti	International Center for Journalists (U.S.); Centre for Freedom of the Media, University of Sheffield (UK); Reuters Institute for the Study of Journalism, University of Oxford, (UK)
Denis Teyssou	Agence France Presse, France
Trisha Meyer	Vrije Universiteit Brussel, Belgium
Sam Gregory	WITNESS, U.S.
Clara Hanot	EU Disinfo Lab, Belgium
Diana Maynard	University of Sheffield, UK

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

ISBN 978-92-3-100403-2

Contents

Figures

Tables

Members of the Working Group on Freedom of Expression and Addressing Disinformation

1. Hessa Al Jaber, chairperson Eshailsat
2. Bocar Ba, SAMENA
3. Moez Chakchouk, Focal Point, UNESCO
4. Piotr Dmochowski-Lipski, EUTELSAT IGO
5. Ramin Guluzade, Republic of Azerbaijan
6. Carlos Manuel Jarque Uribe, America Movil
7. Beeban Kidron, 5Rights Foundation
8. Robert Kirkpatrick, UN Global Pulse
9. Yee Cheong (Dato') Lee ISTIC
10. Adrian Lovett, The Web Foundation
11. Philipp Metzger, DG Swiss Ministry of Communications
12. Paul Mitchell, Microsoft
13. Speranza Ndege, Kenyatta University
14. Patrick Nyirishema, Focal Point, Director-General Rwanda Utilities Regulatory Authority
15. Joanna Rubinstein, World Childhood Foundation
16. Achim Steiner, UNDP (via Minerva Novero-Belec)

Expert oversight group

Fabrício Benevenuto (Federal University of Minas Gerais)

Divina Frau-Meigs (Université Sorbonne Nouvelle - Paris 3).

Cherian George (Hong Kong Baptist University)

Claire Wardle (First Draft)

Herman Wasserman (University of Cape Town)

Acknowledgements:

Guy Berger (UNESCO)

Guilherme Canela De Souza Godoi (UNESCO)

Oscar Castellanos (UNESCO)

Jo Hironaka (UNESCO)

Anna Polomska (ITU)

Cedric Wachholz (UNESCO)

Martin Wickenden (UNESCO)

Joanna Wright (University of Sheffield)

Shanshan Xu (UNESCO)

Additional feedback received from Broadband Commission Working Group member organisations:

UNDP (Minerva Novero Belec, Robert Opp)

UN Global Pulse (Chis Earney)

Foreword

Audrey Azoulay
UNESCO Director General

Dr Hessa al-Jabar
Chairperson Es'hailSat &
Working Group co-chair

As the past few months have shown, online communication is essential in today's world. The Internet plays a key role when it comes to accessing education, culture and quality information. It allows us to work from home while staying in touch with family and friends. Now more than ever, online communication is at the heart of our connected societies.

This greater connectivity goes hand in hand with greater opportunities. It empowers people with information and knowledge, which in turn supports development and democracy. It diversifies language, enables us to do business, and encourages us to appreciate different cultures.

However, online communication can also have a darker side. Incorrect information and misleading messages are surging on digital platforms. The design of algorithms is being exploited through orchestrated behaviours and campaigns, so that content harming human rights and fundamental freedoms is being automatically recommended in news feeds and search results. Not only does this affect our trust in public institutions, it endangers peace and public health.

COVID-19 has brought this issue into sharp relief. As the virus spread across the globe, so too did a flood of rumours and false information. As our report shows, for example, one in four popular YouTube videos on the coronavirus contained misinformation. At a time when scientists around the world are working to develop a vaccine, another study found that more than 1,300 anti-vaccination pages on Facebook had nearly 100 million followers. Yet accurate information is essential to save lives, especially during health crises. In the words of historian Yuval Noah Harari, interviewed in the UNESCO Courier, our best defence against pathogens is information, not isolation. For this reason, balancing freedom of expression and the fight for reliable information has never been so important.

We need to address this issue now – and we need to make sure we have the right tools. This is what this report sets out to do, by identifying and analysing no less than 11 ways of promoting high-quality information. The resulting toolkit includes a wide range of responses, from policy and legislative measures to technological efforts and media and education literacy initiatives.

Professional journalists are central to this toolkit. By identifying and investigating problems, they can track, debunk and deal with lies, while ensuring that legitimate debate does not become a casualty in the fight against falsehoods. UNESCO defends the essential role journalists play in our societies – by encouraging public debate, they help build citizen awareness.

To address these issues, this unique and comprehensive document has been developed under the auspices of the Broadband Commission for Sustainable Development, co-chaired by H.E. President Paul Kagame and Carlos Slim. I would like to thank the Commission's Working Group on Freedom of Expression and Addressing Disinformation for supporting this timely global research.

In today's troubling times, the flood of false and misleading content is exactly what the world does not need. However, as this report shows, by working together, we can defend reliable, high-quality information while advancing freedom of expression. This document is a case in point for the 'digital co-operation' advocated by the United Nations Secretary-General.

We therefore encourage Commissioners and other stakeholders to make full use this report. Together, we can help ensure that broadband for sustainable development achieves its full potential.

Executive Summary

In June 2020, more than 130 United Nations member countries and official observers called on all States to take steps to counter the spread of disinformation, especially during the COVID-19 pandemic (UN Africa Renewal, 2020). They underlined that these responses should:

- Be based on:
  - Freedom of expression,
  - Freedom of the press and promotion of highest ethics and standards of the press,
  - The protection of journalists and other media workers,
- And promote:
  - Media and Information Literacy (MIL).
  - Public trust in science, facts, independent media, state and international institutions.

The need for action against disinformation has also been recognised at the ITU/UNESCO Broadband Commission for Sustainable Development. The Commission created a Working Group on Freedom of Expression and Addressing Disinformation, that in turn commissioned this comprehensive global study in 2019. The research underpinning this study was conducted between September 2019 and July 2020 by an international and interdisciplinary team of researchers.

Balancing Act: *Responding to Disinformation While Defending Freedom of Expression* uses the term 'disinformation' to describe false or misleading content with potentially harmful consequences, irrespective of the underlying intentions or behaviours in producing and circulating such messages. The focus is not on definitions, but on how States, companies, institutions and organisations around the world are responding to this phenomenon, broadly conceived. The work includes a **novel typology of 11 responses**, making holistic sense of the disinformation crisis on an international scale, including during COVID-19. It also provides a **23-step tool** developed to assess disinformation responses, including their impact on freedom of expression (see below).

The research concludes that disinformation cannot be addressed in the absence of freedom of expression concerns, and it explains why actions to combat disinformation should support, and not violate, this right. It also underlines that access to reliable and trustworthy information, such as that produced by critical independent journalism, is a counter to disinformation.

Additionally, the study has produced a framework for capturing the complete disinformation life cycle - from instigation and creation, to the means of propagation, to real-life impact, with reference to: **1**. **Instigators 2**. **Agents 3**. **Messages 4**. **Intermediaries 5**. **Targets**/**Interpreters** - shortened to the acronym '**IAMIT**'.

A series of cascading questions can be asked within the various stages of the life cycle with reference to the actors implicated:

1. Instigators:

Who are the direct and indirect instigators and beneficiaries of the disinformation? What is their relationship to the agent(s) (below)? Why is the disinformation being spread - what is the motivation e.g. political, financial, status boosting, misguided altruism, ideological, etc.? Thus, including, where discernible, if there is intent to harm and intent to mislead.

2. Agents:

Who is operationalising the creation and spread of disinformation? This question raises issues of actor attribution (related to authentic identity), type ('influencer', individual, official, group, company, institution), level of organisation and resourcing, level of automation. Thus behaviours are implicated - such as using techniques like bots, sock puppet networks and false identities.

3. Messages:

What is being spread? Examples include false claims or narratives, decontextualised or fraudulently altered images and videos, deep fakes, etc. Are responses covering categories which implicate disinformation (eg. political/electoral content)? What constitutes potentially harmful, harmful, and imminently harmful messaging? How is false or misleading content mixed with other kinds of content - like truthful content, hateful content, entertainment and opinion? How is the realm of unknowns being exploited by disinformation tactics? Are messages seeking to divert from, and/or discredit, truthful content and actors engaged in seeking truth (e.g. journalists and scientists)?

4. Intermediaries:

- Which sites/online services and news media is the disinformation spreading on? To what extent is it jumping across intermediaries, for example starting on the 'dark web' and ending up registering in mainstream media?

- How is it spreading? What algorithmic and policy features of the intermediary site/app/network and its business model are being exploited? Do responses seek to address algorithmic bias that can favour disinformation? Also, is there evidence of coordinated behaviour (including inauthentic behaviour) exploiting vulnerabilities, in order to make it appear that specific content is popular (even viral) when in fact it may have earned this reach through deliberately gaming the algorithms?

- Are intermediaries acting in sufficiently accountable and transparent ways and implementing necessary and proportionate actions to limit the spread of disinformation?

5. Targets/Interpreters:

- Who is affected? Are the targets individuals, journalists and scientists, systems (e.g. electoral processes; public health; international norms); communities; institutions (like research centres); or organisations (including news media);

- What is their online response and/or real-life action? This question covers responses such as inaction, sharing as de facto endorsement, liking, or sharing to debunk disinformation. Is there uncritical news reporting (which then risks converting the role of a complicit journalist/news organisation from target into a disinformation agent)?

- Responses identifying what messages count as disinformation, investigating the instigators and agents, identifying the intentions and targets;

- Responses aimed at circumscribing and countering the agents and instigators of disinformation campaigns;

- Responses aimed at curtailing the production and distribution of disinformation and related behaviours, implemented particularly by intermediaries and media;

- Responses aimed at supporting the targets/interpreters of disinformation campaigns.

Eleven response types are then identified and assessed under four umbrella categories:

1. **Identification responses** (aimed at identifying, debunking, and exposing disinformation)

   i. Monitoring and fact-checking

   ii. Investigative

2. **Responses aimed at producers and distributors through altering the environment that governs and shapes their behaviour**

   iii. Legislative, pre-legislative, and policy responses

   iv. National and international counter disinformation campaigns

   v. Electoral responses

3. **Responses aimed at production and distribution mechanisms** (pertaining to the policies and practices of institutions mediating content)

   vi. Curatorial responses

   vii. Technical and algorithmic responses

   viii. Demonetisation responses

4. **Responses aimed at the target audiences of disinformation campaigns** (aimed at supporting the potential 'victims' of disinformation)

   ix. Normative and ethical

   x. Educational

   xi. Empowerment and credibility labelling responses

These responses to disinformation are shown to often be complementary to each other. For example, in many cases, investigations by journalists have exposed online disinformation that had remained undetected (or unrecognised) in the monitoring and fact-checking organised by the internet communication companies. In other words, actions taken by the companies alone to stop transmission of disinformation depend in part on the work of investigation by other actors. Similarly, even if some efforts help cut the supply and transmission of disinformation, there is still a need to empower the targets against that content which does reach them, and thereby at least aid prevention of viral recirculation.

The study also finds that there are cases where one type of response can work against another. An example is an over-emphasis on responses through top-down regulation at the expense of bottom-up empowerment. Further, there is the phenomenon of catching journalists in nets set for disinformation agents through the criminalisation of the publication or distribution of false information (e.g. via 'fake news' laws). This works directly against the role of independent, critical journalism as a counter to disinformation. A similar example exists in cases of internet communications companies not removing disinformation-laden attacks on journalists on the grounds of 'free speech'. In this way, a very particular understanding of expression undermines press freedom and journalism safety, and therefore journalism's service against disinformation.

These illustrations signal that different interventions need to be aligned, rather than going in separate directions. Accordingly, this study calls for multistakeholder consultation and cooperation in the fight against disinformation. This aligns with UNESCO's model of Internet Universality, which upholds the principle of multistakeholder governance in digital issues.

The study further recognises that a multi-faceted approach is needed - including addressing socio-economic drivers of disinformation, through rebuilding the social contract and public trust in democratic institutions, promotion of social cohesion, particularly in highly polarised societies, and addressing business models that thrive on paid disinformation content such as advertising that crosses the line, through to fraudulent content masquerading as legitimate news or factually-grounded opinion.

For all those seeking to intervene against disinformation, this study urges that each actor include systematic monitoring and evaluations within their response activities. These should cover effectiveness, as well as impacts on the right to freedom of expression and access to information, including on the right to privacy.

The findings also underline the need for increased transparency and proactive disclosure across all 11 kinds of responses to disinformation. This aligns with the spirit of Sustainable Development Goal target 16.10 which calls for public access to information and fundamental freedoms.

Among other measures, the research encourages the broadband community and donors to invest further in independent fact-checking, critical professional journalism, media development and Media and Information Literacy (MIL), especially through educational interventions targeting children, young people, older citizens, and vulnerable groups. It also calls for actors to promote privacy-preserving, equitable access to key data from internet communications companies, to enable independent analysis into the incidence, spread and impact of online disinformation on citizens around the word, and especially in the context of elections, public health, and natural disasters.

In addition to these overarching proposals, the study addresses key stakeholder groups, making a set of recommendations for action in each case. Among these, the following recommendations are highlighted here:

Intergovernmental and other international organisations, **as appropriate**, **could**:

- Increase technical assistance to Member States at their request in order to help develop regulatory frameworks and policies, in line with international freedom of expression and privacy standards, to address disinformation. This could involve encouraging the uptake of the 23-step disinformation response assessment tool developed for this study (see below).

- Particularly in the case of UNESCO with its mandate on freedom of expression, step up the work being done on disinformation in partnership with other UN organisations and the range of actors engaged in this space.

Individual states could:

- Actively reject the practice of disinformation peddling, including making a commitment not to engage in public opinion manipulation either directly or indirectly - for example via 'influence operations' produced by third party operators such as'dark propaganda' public relations (PR) firms.

- Review and adapt their responses to disinformation, using the 23-step tool for assessing law and policy developed as an output of this study, with a view to conformity with international human rights standards (notably freedom of expression, including access to information, as well as privacy rights), and at the same time making provision for monitoring and evaluation of their responses.

- Increase transparency and proactive disclosure of official information and data, and monitor this performance in line with the right to information and SDG indicator 16.10.2 that assesses the adoption and implementation of constitutional, statutory and/or policy guarantees for public access to information.

Political parties and other political actors could:

- Speak out about the dangers of political actors as sources and amplifiers of disinformation and work to improve the quality of the information ecosystem and increase trust in democratic institutions.

- Refrain from using disinformation tactics in political campaigning, including the use of covert tools of public opinion manipulation and 'dark propaganda' public relations firms.

Electoral regulatory bodies and national authorities could:

- Improve transparency of all election advertising by political parties, candidates, and affiliated organisations through requiring comprehensive and openly available advertising databases and disclosure of spending by political parties and support groups.

- Work with journalists and researchers in fact-checking and investigations around electoral disinformation networks and producers of 'dark propaganda'.

Law enforcement agencies and the judiciary could:

- Ensure that law enforcement officers are aware of freedom of expression and privacy rights, including protections afforded to journalists who publish verifiable information in the public interest, and avoid arbitrary actions in connection with any laws criminalising disinformation.

- For judges and other judicial actors: Pay special attention when reviewing laws and cases related to addressing measures to fight disinformation, such as criminalisation, in order to help guarantee that international standards on freedom of expression and privacy are fully respected within those measures.

Internet communications companies could:

- Work together in a human rights frame, to deal with cross-platform disinformation, in order to improve technological abilities to detect and curtail false and misleading content more effectively and share data about this.

- Develop curatorial responses to ensure that users can easily access journalism as verifiable information shared in the public interest, prioritising news organisations that practice critical, ethical independent journalism.

- Recognise that if health disinformation and misinformation can be quickly dealt with in a pandemic on the basis that it poses a serious risk to public health, action is also needed against political disinformation - especially at the intersection of hate speech – when it, too, can be life-threatening. The same applies to disinformation related to climate change.

- Recognise that press freedom and journalism safety are critical components of the internationally enshrined right of freedom of expression, meaning that online violence targeting journalists (a frequent feature of disinformation campaigns) cannot be tolerated.

- Apply fact-checking to all political content (including advertising, fact-based opinion, and 'direct speech') published by politicians, political parties, their affiliates, and other political actors.

The study also addresses recommendations to other stakeholder groups such as news media, civil society organisations, advertising brokers, and researchers.

In totality, this research affirms that freedom of expression, access to information and critical, independent journalism - supported by open and affordable internet access - are not only fundamental human rights, but should be treasured as essential tools in the arsenal to combat disinformation - whether connected to a pandemic, elections, climate

change or social issues. This timely study serves as a call to all stakeholders to uphold these international norms which, along with the UN's sustainable development goals, are under significant threat from disinformation.

It cautions that the fight against disinformation is not a call to suppress the pluralism of information and opinion, nor to suppress vibrant policy debate. It is a fight for facts, because without widely available evidence-based information, access to reliable, credible, independently verifiable information that supports democracy and helps avert worsening the impacts of crises like pandemics will not be possible.

The 'cures' for disinformation should not exacerbate the 'disease', nor create challenges worse than the problem itself. But working together, those actors involved in implementing initiatives within the 11 response types covered in this study, can ensure that their actions are transparent, gender-sensitive, human-rights compliant, systematically evaluated ... and optimally effective.

Assessment Framework for Disinformation Responses

The study offers a Freedom of Expression Assessment Framework for Disinformation Responses to assist UNESCO Member States and other institutions to formulate legislative, regulatory and policy responses to counter disinformation in a manner that supports freedom of expression. The tool includes 23 reference points to enable assessment of responses in accordance with international human rights norms, paying additional attention to access to information and privacy rights.

1. Have responses been the subject of multi-stakeholder engagement and input (especially with civil society organisations, specialist researchers, and press freedom experts) prior to formulation and implementation? In the case of legislative responses, has there been appropriate opportunity for deliberation prior to adoption, and can there be independent review?

2. Do the responses clearly and transparently identify the specific problems to be addressed (such as individual recklessness or fraudulent activity; the functioning of internet communications companies and media organisations; practices by officials or foreign actors that impact negatively on e.g. public health and safety, electoral integrity and climate change mitigation, etc)?

3. Do responses include an impact assessment as regards consequences for international human rights frameworks that support freedom of expression, press freedom, access to information or privacy?

4. Do the responses impinge on or limit freedom of expression, privacy and access to information rights? If so, and the circumstances triggering the response are considered appropriate for such intervention (e.g. the COVID-19 pandemic), is the interference with such rights narrowly-defined, necessary, proportionate and time limited?

5. Does a given response restrict or risk acts of journalism such as reporting, publishing, and confidentiality of source communications, and does it limit the right of access to public interest information? Responses in this category could include: 'fake news' laws; restrictions on freedom of movement and access to information in general, and as applied to a given topic (eg. health statistics, public expenditures); communications interception and targeted or mass surveillance;

data retention and handover. If these measures do impinge on these journalistic functions or on accountability of duty-bearers to rights-holders in general, refer to point 4. above.

6. If a given response does limit any of the rights outlined in 4., does it provide exemptions for acts of journalism?

7. Are responses (eg. educational, normative, legal, etc.) considered together and holistically in terms of their different roles, complementarities and possible contradictions?

8. Are responses primarily restrictive (eg. legal limits on electoral disinformation), or there is an appropriate balance with enabling and empowering measures (eg. increased voter education and Media and Information Literacy)?

9. While the impacts of disinformation and misinformation can be equally serious, do the responses recognise the difference in motivation between those actors involved in deliberate falsehood (disinformation) and those implicated in unwitting falsehood (misinformation), and are actions tailored accordingly?

10. Do the responses conflate or equate disinformation content with hate speech content (even though international standards justify strong interventions to limit the latter, while falsehoods are not per se excluded from freedom of expression)?

11. Are journalists, political actors and human rights defenders able to receive effective judicial protection from disinformation and/or hateful content which incites hostility, violence and discrimination, and is aimed at intimidating them?

12. Do legal responses come with guidance and training for implementation by law enforcement, prosecutors and judges, concerning the need to protect the core right of freedom of expression and the implications of restricting this right?

13. Is the response able to be transparently assessed, and is there a process to systematically monitor and evaluate the freedom of expression impacts?

14. Are the responses the subject of oversight and accountability measures, including review and accountability systems (such as reports to the public, parliamentarians, specific stakeholders)?

15. Is a given response able to be appealed or rolled-back if it is found that any benefits are outweighed by negative impacts on freedom of expression, access to information and privacy rights (which are themselves antidotes to disinformation)?

16. Are measures relating to internet communications companies developed with due regard to multi-stakeholder engagement and in the interests of promoting transparency and accountability, while avoiding privatisation of censorship?

17. Is there assessment (informed by expert advice) of both the potential and the limits of technological responses which deal with disinformation (while keeping freedom of expression and privacy intact)? Are there unrealistic expectations concerning the role of technology?

18. Are civil society actors (including NGOs, researchers, and the news media) engaged as autonomous partners in regard to combatting disinformation?

19. Do responses support the production, supply and circulation of information - including local and multilingual information - as a credible alternative to disinformation? Examples could be subsidies for investigative journalism into disinformation, support for community radio and minority-language media.

20. Do the responses include support for institutions (e.g. public service messaging and announcements; schools) to enable counter-disinformation work? This could include interventions such as investment in projects and programmes specifically designed to help 'inoculate' broad communities against disinformation through Media and Information Literacy (MIL) programmes.

21. Do the responses maximise the openness and availability of data held by state authorities, with due regard to personal privacy protections, as part of the right to information and official action aimed at pre-empting rumour and enabling research and reportage that is rooted in facts?

22. Are the responses gender-sensitive and mindful of particular vulnerabilities (e.g. youth, the elderly) relevant to disinformation exposure, distribution and impacts?

23. If the response measures are introduced to respond to an urgent problem, or designed for short term impact (e.g. time sensitive interventions connected to elections) are they accompanied by initiatives, programmes or campaigns designed to effect and embed change in the medium to long term?

Introduction

1

Authors: Kalina Bontcheva and Julie Posetti

This global study seeks to map and deepen understanding of diverse international responses to disinformation, along with the impacts of counter-disinformation measures on the right to freedom of opinion and expression, as described in Article 19 of the United Nations' Universal Declaration of Human Rights[1]:

> *"Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."*

Freedom of expression rights, including press freedom and the right to access information, are upheld in tandem with privacy rights, which are also enshrined in international human rights law. So, where relevant, this study also touches on online privacy and dignity issues. Further, it situates the problem of disinformation in the context of the enabling role of the internet - especially the social web - in both improving access to information, and as a disinformation vector. It discusses in detail the potential for responses to disinformation to curb freedom of expression and suggests ways to avoid such impacts.

Although many studies and policy papers on disinformation have already been published by governments, international organisations, academics, and independent think tanks, this study offers novel contributions through its development of a systematic typology of the range of responses to disinformation which is applied internationally:

1. Addressing the entire spectrum of disinformation responses, rather than e.g. just educational or legal or technological responses;
2. Categorising responses according to the target of the intervention, rather than in terms of the means used or the actors involved;
3. Assessing responses in terms of key assumptions and significance from a freedom of expression point of view;
4. Representation of geographically diverse issues, cases and responses, including an emphasis on the Global South;
5. Providing an overview of disinformation responses aimed at 'flattening the curve' of the COVID-19 'disinfodemic' (Posetti & Bontcheva 2020a; Posetti & Bontcheva 2020b).

There are diverse definitions applied to false and misleading information, but for the purposes of this study the term disinformation is used throughout to broadly refer to content that is false and has potentially damaging impacts - for example, on the health and safety of individuals and the functionality of democracy. For many analysts, the intent of the agent producing or sharing the inaccurate content can also differentiate disinformation (deliberate falsehood) from misinformation (unconscious falsehood). This study accepts the role of such a distinction, which also implicates different types of remedies. Nevertheless, the impact of the false content, irrespective of intentions, can be the same. It is this focus on the potentially damaging effects of fabricated and misleading content, rather than the motivation for its creation and dissemination, that explains the broad use of the term disinformation here as the umbrella term - irrespective of intentionality or underlying behaviour in spreading such messages. This rationale is further explained in section 1.2 on definitions below.

[1] https://www.un.org/en/universal-declaration-human-rights/

Disinformation (as opposed to verifiable information) can cause harm since it may serve to confuse or manipulate citizens, create distrust in international norms, institutions or democratically agreed strategies, disrupt elections, or paint a false picture about key challenges such as climate change. It can also be deadly, as the COVID-19 'disinfodemic' has illustrated (Posetti and Bontcheva 2020a; Posetti and Bontcheva 2020b). Disinformation is typically organised by both state or non-state actors, including individuals and organised groups. It is created, spread and amplified both organically, by people who believe it, and artificially through campaigns that make use of technology such as bots and recommender algorithms. It is crafted to exploit cognitive biases such as attentional and confirmation biases, while using astroturfing[2] techniques to stimulate what is known as the 'bandwagon effect' (Schmitt-Beck, 2008), creating the impression of widely shared beliefs around a particular issue or item. Frequently, disinformation campaigns aim to target, discredit, and silence those who produce verified information or hold opposing views, including politicians, journalists, human rights campaigners, scientists, and others. Many disinformation agents carry out campaigns that are also networked across different platforms and combined with threats, intimidation and disruptive tactics.

In particular, disinformation negatively impacts citizens' rights to privacy, freedom of expression and access to information. In turn, however, many efforts to tackle online disinformation can also interfere with these fundamental human rights, as discussed throughout this report. Tools, measures and policies to address the disinformation problem therefore need to ensure that the rights of citizens are protected, and that their interests are represented. This means taking an approach that acknowledges how the issues affect stakeholders such as journalistic actors, civil society organisations, and the internet communications companies.[3] Frequently, however, these rights and interests are in tension in the struggle to identify, curtail and counter disinformation. For example, what's the interplay between content moderation, freedom of speech, and algorithmic amplification of misinformation?

Under human rights law, expression of false content - like other expression - is protected, with some exceptions. For example, under the International Covenant on Civil and Political Rights, certain forms of hate speech, incitement to violence, and speech that threatens human life (including dangerous health disinformation) can attract legitimate restrictions for reasons such as the protection of other human rights, or for public health purposes. Nevertheless, inasmuch as speech does not reach this threshold of legitimate restriction, people have a right to express ill-founded opinions and make non-factual and unsubstantiated statements - ranging from claims that "The earth is flat" to opinion like "The unusually cold weather we are experiencing means that global warming must be a myth in my view" - including on social media (Allan, 2018). On the other hand, falsehoods designed to defraud people financially, defame a person's reputation, or suppress voter turn-out, may be fairly penalised under criminal or civil law in many cases. All this makes tackling disinformation even more complex from the point of view of freedom of expression.

2 'Astroturfing' is a term derived from a brand of fake grass used to carpet outdoor surfaces to create the impression that it is natural grass cover. In the context of disinformation, it involves seeding and spreading false information, targeting audiences and journalists with an intention to redirect or mislead them, particularly in the form of 'evidence' of faux popular support for a person, idea or policy. See also Technopedia definition: https://www.techopedia.com/definition/13920/astroturfing

3 Throughout this report, the term 'internet communications companies' is used to refer to large companies in the sphere of search engines, social media sites and messaging apps. This avoids the practice of referring to these companies generically as 'the platforms' in order to underline their diversity, and because they are not neutral or passive technological infrastructural services but institutions with interests, obligations and configurations that have significant bearing on information, disinformation and communications.

Contemporary expression is closely intertwined with the combination of information technologies and internet communications companies which, coupled with growing broadband access, enable the instantaneous dissemination of information within global networks that are accessible to billions of people. This facilitates freedom of expression and potentially opens up a far wider range of viewpoints and information sources to citizens than ever before. In a world divided between information-rich and information-poor, this is seen as a boon to people who have previously been uninformed. Conversely, however, these tools of freedom of expression have been increasingly weaponised by actors seeking to manipulate public opinion by inserting and amplifying false and misleading content within the online information ecosystem.

The increasing availability of information, coupled with the potential for more diverse news diets, could widen the range of ideas to which people are exposed. Within the vast sea that is the contemporary information ecosystem, there are credible information providers like those journalism producers who do live up to the standards of independent professionalism, independent research institutes, other producers of reliable public interest information (e.g. reputable health advice providers), and well-informed commentators. But there is also a mass of other players with different standards of truthfulness, diverse ethics and varying motives creating a powerful rip current within this sea. Consequently, citizens can feel overwhelmed by the flood of content they are exposed to online, and they can come to rely on spurious sources that appeal to their biases and reinforce their pre-existing beliefs or identities. As a result, in place of being uninformed, they may become actively disinformed, or indirectly misinformed.

Recent research has demonstrated that disinformation affects different countries to various extents (Humprecht, Esser & Van Aelst, 2020). Increased ideological segregation and political polarisation are some of the key drivers behind the elevated production and spread of online disinformation in some countries (Humprecht, Esser & Van Aelst, 2020). By contrast, other research indicates that digital information consumption can lead to exposure to a broader range of information sources, although it does not necessarily follow that the content is itself more diverse, nor that the beliefs held are therefore diversified. However, repetitious exposure to falsehoods is known to reduce resistance to disinformation, as does exposure to high levels of populist communication (Humprecht, Esser & Van Aelst, 2020).

Conversely, resilience to disinformation is higher in countries where trust in news media is high and public service media provision is strong. Moreover, low public trust in news media and democratic institutions can lead to highly selective information consumption through online echo chambers that amplify disinformation and deepen polarisation.

Consequently, there is an urgent need to not only address disinformation, but also to take steps towards rebuilding the social contract and public trust in agreed international norms and standards: strengthen democratic institutions; promote social cohesion particularly in highly divided societies; and engage dialogue-building tactics to address entrenched groups and actors online.

This is why it is imperative to examine the diverse responses to disinformation globally, and to develop frameworks to help understand and assess these responses through a freedom of expression lens. That is the primary work of this study, research for which was conducted between September 2019 and July 2020.

Before this work turns to deciphering and dissecting these dimensions, it is necessary to outline the parameters for the research, explain the key terms used, and consider some examples of online disinformation, along with their relationship to propaganda, misinformation, and hate speech.

1.1 Techniques of online disinformation

The ubiquitous presence of online disinformation poses serious questions about the role of search, social media and social messaging and the internet more widely in contemporary democracies. Examples of digital disinformation abound, ranging from election interference to medical disinformation (e.g. vaccination[4]; coronavirus[5]) and these frequently involve threats of physical harm, privacy risks, and reputational damage to individuals and public health.

While disinformation is often studied in regard to Twitter, Facebook, and YouTube, it also exists on many other social platforms (e.g. Reddit, Instagram, TikTok, 4chan, Pinterest[6]), messaging apps (e.g. WhatsApp, Telegram, SnapChat, and iMessage), and internet search engines (e.g. Google). There are also dedicated disinformation sites (e.g. Infowars, Q-anon). Additionally, other actors and intermediaries (e.g. ISPs, cloud computing providers) will also be referenced here where relevant. The study, while comprehensive at the time of writing, also acknowledges the need to continue research into emerging disinformation mechanisms and new and rapidly evolving social platforms, including those received or perceived mainly as entertainment and social spaces (e.g. TikTok) and not as platforms for political and other purposes.

While political actors and States are often significant producers and conduits of disinformation (Brennan et al 2020; Billings 2020; Bradshaw & Howard 2019), the emphasis of this report is not on disinformation sources and actors, but on the responses to disinformation found across the world. Among these responses, States and political actors have a critical role to play in stemming disinformation at the source - including within their own 'houses'. Their relevance is especially assessed in regard to responses concerning counter-messaging, legislation and policy, elections and normative interventions.

Many mainstream news producers - online and offline - struggle to remain a reference point for those seeking trustworthy information within this wider ecology of communications. Through weak standards of verification, manipulation by outside actors, and even complicity (e.g hyper-partisan media), news outlets have also become vectors for disinformation in certain cases.

Nevertheless, the legitimating and agenda-setting public role of critical independent news media also makes them prime targets for purveyors of disinformation. In the case of orchestrated disinformation campaigns, attacks are frequently deployed against legitimate and authoritative information sources - such as credible news media and journalists - through hacking, disruption, and other tactics of intimidation and surveillance, with a view to a holistic strategy for advancing disinformation and wider objectives. Many orchestrated disinformation campaigns are State-initiated and/or connected to political and geopolitical actors, and this is relevant to understanding State roles in the responses to disinformation. However, the primary purpose of this report is to unpack the diverse modalities of response to the global disinformation crisis, rather than assessing the initiators and agents and their motives.

4 https://firstdraftnews.org/long-form-article/first-draft-case-study-understanding-the-impact-of-polio-vaccine-disinformation-in-pakistan/

5 https://www.poynter.org/fact-checking/2020/coronavirus-fact-checkers-from-30-countries-are-fighting-3-waves-of-misinformation/%20

6 https://medium.com/dfrlab/trudeaus-and-trudeaunts-memes-have-an-impact-during-canadian-elections-4c842574dedc

Within the ecosystem, disinformation knows no boundaries, but rather permeates multiple communication channels by design, or through redistribution and amplification fuelled by the architectures of interconnecting peer-to-peer and friend-to-friend networks.

With respect to types of content, three main disinformation formats have been identified for this study, based on the modality of the content (e.g. text, image, video, audio, mixed) and the way it has been constructed or manipulated:

1. **Emotive narrative constructs and memes**: False claims and textual narratives[7] which often (but not always) mix strong emotional language, lies and/or incomplete information, and personal opinions, along with elements of truth. These formats are particularly hard to uncover on closed messaging apps and they are applied to a range of content from fabricated 'news' to problematic political advertising.

    - **False/misleading narratives** emulating formats like news writing or documentary, and which typically mix false textual claims or incomplete information with personal opinions, along with images and/or video and/or audio, which themselves could be inauthentic, manipulated, or decontextualised. Appropriated content from other websites is sometimes used to create a misleading overall impression of being a neutral news-aggregator.

    - **Emotional narratives** with strong personal opinions, images and/or videos and audio, which may be inauthentic, manipulated, or decontextualised, and which also seek to dictate interpretations of particular information at hand, e.g. minimising its significance, smearing the source.

2. **Fraudulently altered**, **fabricated**, **or decontextualised images**, **videos[8] and synthetic audio[9]** used to create confusion and generalised distrust and/or evoke strong emotions through viral memes or false stories. These are also applied to a wide range of content from political propaganda to false advertising. Among these techniques we can distinguish:

    - **Decontextualised images and videos** that are unchanged or almost unchanged with high level of similarity, and often including copies that are used for clickbait purposes;

    - **Altered decontextualised audio**, **images and videos that** are cut in length to one or several fragments of the original audio or video, or changed to remove a timestamp in CCTV camera footage, for example. These are also called 'shallow fakes';

    - **Staged videos** e.g. produced on purpose by a video production company;

    - **Tampered images and videos** that are created with the help of editing software to remove, hide, duplicate or add some visual or audio content;

7 A database of over 6,000 fact-checked false claims and narratives on COVID-19 from over 60 countries: https://www.poynter.org/coronavirusfactsalliance/

8 Decontextualised images or videos are pre-existing, authentic content, which is re-purposed as part of a false narrative to spread disinformation, e.g. an old video of people praying was used in a far-right tweet claiming that Muslims are flouting social distancing rules.

9 See definition below

- **Computer-Generated Imagery** (**CGI**) including deepfakes (false images/videos generated by artificial intelligence) that are entirely computer-generated, or mixed with a blend of pre-existing image/footage/audio.

- **Synthetic audio**: Speech synthesis, where advanced software is used to create a model of someone's voice is a relatively new branch of deepfakes. This involves replicating a voice, which can verbalise text with the same cadence and intonation as the impersonated target. Some technologies (e.g. Modulate.ai) allow users to create completely synthetic voices that are able to mimic any gender or age. (Centre for Data Ethics and Innovation 2019)

3. **Fabricated websites and polluted datasets**, including false sources, manipulated datasets, and fake government or company websites (Trend Micro, 2020). This category also includes websites using names that make them sound like news-media and which publish seemingly plausible information in the genre of news stories, e.g. reporting bogus cases of COVID-19 (Thompson, 2020).

These different disinformation modalities are harnessed in a range of potentially harmful practices, including but not limited to:

- State-sponsored disinformation campaigns;
- (Anti-)Government /Other political propaganda;
- Political leaders generating and amplifying false and misleading content
- Clickbait[10];
- False or misleading advertisements e.g. connected to politics, job adverts;
- Impersonation of authoritative media, fact-checking organisations, people, governments (false websites and/or social media accounts, bots);
- Astroturfing campaigns;
- Fake products and reviews
- Anti-vaccine, coronavirus, and other other health, medical and well-being related misinformation;
- Gaslighting[11];
- Inauthentic identities and behaviours;

Overt satire and parody are excluded from this list of communication practices, even though in some instances these may have the potential to mislead and thus cause harm to citizens who lack sufficient Media and Information Literacy (MIL) competencies to distinguish them. Satire and parody can, in fact, serve as effective counters to disinformation by highlighting the absurd elements of disinformation (and those who create and disseminate it) in effective and engaging ways. However, these communications practices should not generally be treated as constituting disinformation.

[10] A post designed to provoke emotional response in its readers (e.g. anger, compassion, sadness, fear), and thus causes the user to stimulate further engagement (i.e. 'click') by following the link to the webpage, which in turn generates ad views and revenues for the website owner. The defining characteristic of clickbait is that it fails to deliver on the headline, meaning the 'clicker' has taken the bait but the article will not fulfil expectations.

[11] A form of psychological manipulation: https://dictionary.cambridge.org/dictionary/english/gaslighting

1.2 Definitions and scope

There are many different and somewhat contradictory definitions of disinformation, and whether and how it relates to misinformation. The conceptualisations generally share the trait of falsity as an essential criterion, with the result that the terms mis- and dis-information are often used synonymously and interchangeably (e.g. in Alaphilippe, Bontcheva et al., 2018b).

For its part, the Oxford English Dictionary[12] (OED) appears to distinguish the labels on the basis of one being linked to the intention to deceive, and the other the intention to mislead (although it is not clear how these objectives differ):

- **Misinformation**: False or inaccurate information, especially that which is deliberately intended to deceive.
- **Disinformation**: False information which is intended to mislead, especially propaganda issued by a government organisation to a rival power or the media.

This definition also links one of the terms (disinformation) to a particular actor (governmental), which would seem to suggest a narrowing of the scope of its remit. Others have defined disinformation specifically in the context of elections, as "*content deliberately created with the intent to disrupt electoral processes*" (Giglietto et al., 2016). This definition is likewise too narrow for the wider variety of disinformation considered in this study.

A further perspective is evident in the recommendations of a report produced by the EU High Level Expert Group on Fake News and Online Disinformation, which includes references to possible for-profit disinformation as part of what is covered by the term:

> *"Disinformation....includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or [generate] profit. (Buning et al., 2018)."*

But intention to profit is a potentially limiting factor. For example, satire is created for profit by television producers and satirical magazines, and it would be problematic to include this communications practice as disinformation per se.

The widely-adopted *information disorder* theoretical framework (Wardle, 2017a; Wardle & Derakhshan, 2017) distinguishes mis- and disinformation as follows:

- **Misinformation**: false information that is shared inadvertently, without meaning to cause harm.
- **Disinformation**: intending to cause harm, by deliberately sharing false information

The underlying criteria in this framework could be represented as such:

[12] https://www.lexico.com/definition/misinformation

	Awareness of falsity	Underlying intent
Disinformation	Aware	"Bad"
Misinformation	Unaware ("inadvertent")	"Good / neutral"

These definitions broadly align with those in the Cambridge English Dictionary[13], where disinformation is defined as having intention to deceive, whereas misinformation is more ambiguous.

Most definitions share the feature of intentionality regarding harm (implicit in the OED semantics is that both deception and attempts to mislead are negative practices).

At the same time, operationalising a distinction based on intention (and awareness of falsity) is complicated by the fact that the motivation and knowledge of the information source or amplifier may often not be easily discernible, not only by algorithms, but also by human receivers (Jack, 2017; Zubiaga et al., 2016). There is also a risk of a Manichean assumption about who is "a bad actor", which can greatly over-simplify the situation, and entail highly subjective or problematically partisan interpretations of what and whose interests are intended to be harmed.

What this highlights is the challenge of placing intentionality and awareness of falsehood at the core of the definition of what should count as disinformation, in the face of a wider phenomenon of false or misleading content. This partially explains why some writers (eg. Francois, 2019) approach the issue not by intention (or agent awareness) in the first instance but instead by putting attention on visible behaviours such as coordinated operations involving bots (which may suggest harmful intention and awareness of falsity at play). It is the case that orchestrated behaviours (including by inauthentic actors) can signal false content, yet potentially harmful falsehoods can also spread without special amplification, and they all too often originate from authentic actors like celebrities and politicians, as shown in research (e.g. Brennen et al 2020; Satariano & Tsang 2019). At the same time, truthful content may be circulated through various behaviours and actors as part of an information or counter-disinformation campaign, which is distinct from what is recognised in regard to decontextualised or falsely contextualised content in the term 'malinformation' by Wardle & Derakshan (2017). For these reasons, it would be limiting to reduce the scope of this study to treating disinformation as if the phenomenon was defined essentially by behaviours (as much as they may often be a flag for problems).

For its part, because this study seeks to cover the wide range of responses in play around the world, it avoids a narrow approach to defining disinformation. Accordingly, it uses the term disinformation generically to describe false or misleading content that can cause specific harm - irrespective of motivations, awareness, or behaviours. Such harm may be, for example, damage to democracy, health, minority and disadvantaged communities, climate challenges, and freedom of expression. Here, therefore, the operational approach to what constitutes disinformation (and hence responses to the phenomenon) are the characteristics of falsity and potentially negative impact on targets, rather than the intentionality, awareness or behaviours of its producers(s) or distributor(s). Further, if we understand misinformation in the narrow sense of inadvertent sharing without intent to cause harm, it is evident that the content at hand often owes its origin to others' deliberate acts of disinforming citizens with harmful intent. Acknowledging this 'source' of

13 https://dictionary.cambridge.org/dictionary/english/misinformation

much of the harm is a strong reason for adopting 'disinformation' as the generic term in this study, rather than 'misinformation'.

This approach is not to be reductionist in the sense of categorising all content as either potentially harmful disinformation or (true) information which does not inflict harm. Opinion reflecting values and attitudes is one example of content that cannot be classed as true or false. Science and policy, as another example, are matters in process which evolve over time and may, at least initially, resist a binary assessment. For its part, disinformation, by its nature, claims as 'true' not only falsehoods but also often what is the category of the unknown, while frequently seeking to discredit as 'false' that content that has been definitively proven to be true - such as the overwhelming scientific consensus on climate change. It is because of the existence of genuine grey areas, that there are risks in any steps taken to counter disinformation which disregard the large realm of unknowns which exist between proven truth and demonstrated falsehoods. Such measures can stifle legitimate debate and other forms of expression which are needed to help assess the veracity of particular content over time.

The use of disinformation as a generic term applied to assess responses to false content does not preclude recognition that these responses may vary according to the diverse motivations (financial, political, ideological, personal status, etc) or behaviours of the implicated disinformational instigators and actors. For example, education is a partial remedy for *misinformation* (when understood to refer to unwitting creation or circulation of falsehoods without ill intent or awareness that the content is not true), while regulation to stop money-making from scams is one of the ways to reduce the supply of *disinformation* (using the latter term here in the narrow sense to refer to conscious and deliberate lying). Deliberate distortions and deception may be more prevalent in political and electoral contexts, while misinformation (in the narrow sense) is possibly a greater factor in the case of anti-vaccination content. The underlying theory of change entailed within a given response, is thus often linked to assumptions about intent and related behaviours. Nevertheless, especially in the context of elections, referenda, and pandemics like COVID-19, the harmful impact of false content, irrespective of intentions, and irrespective of the range of behaviours underlying them, is potentially the same. People are disempowered and serious impacts can result. So, interventions need to be appropriately calibrated.

Given the remit of this study, it makes sense for the semantic framing to use the term 'disinformation' as a meta-label to cover falsehoods (encompassing misleading messages) within content and which are associated with potential societal harm (such as negative impacts on human rights, public health and sustainable development). It is this that enables the wide-ranging unpacking of the responses to disinformation underway worldwide. The intent, therefore, is not to produce yet another definition of what disinformation is, but to provide for a broad umbrella conceptualisation of the field under examination and analysis. On this broad foundation, the research that follows takes care to signal, where appropriate, how various stakeholders responding to disinformation interpret the phenomenon, implicitly or explicitly - in regard to the particular type of response under discussion.

Adopting such an approach, this study is able to show how the complex disinformation phenomenon is being met with varying responses around the world, and the bearing that these responses have on freedom of expression and sustainable development. At the same time, it is worth highlighting how this study perceives what disinformation is not. Accordingly, disinformation should not be reduced to falsity with potential harm only in news content (as is implied in the label "fake news") and, as elaborated below, it should also not be conflated with propaganda or hate speech.

1.3 Conceptualising the life-cycle of disinformation

In order to fully understand the responses that seek to counter online disinformation effectively, it is necessary to focus not only on the message itself and its veracity, but also to investigate all aspects of the disinformation lifecycle, including its spread and effects on the target recipients.

One conceptual framework is called the 'ABC' framework, distinguishing between Actors, Behaviour and Content. This attempts to give attention to 'manipulative' actors who engage knowingly in disinformation, to inauthentic and deceptive network behaviour such as in information operations, and to content that spreads falsehoods (using manipulated media formats), or that which may be factual but is inflammatory (Francois, 2019; Annenberg Public Policy Center, 2020). The motivation here is to encourage responses to avoid acting against content that may be 'odious' but which should qualify as protected speech in a democratic society. It therefore points attention to the issue of whether responses should better focus on A and B more than C.

'AMI' is another conceptual approach (Wardle & Derakhshan, 2017), which distinguishes between:

- the **Agents**, i.e. the authors or distributors of disinformation and their motivations;
- the **Message**, i.e. the false and/or manipulated content that is being spread; the way it is expressed, and the techniques used to enhance its credibility;
- the **Interpreters** (or **Targets**), i.e. those targeted by the disinformation campaign and the effects on their beliefs and actions.

In this study, these two frameworks are adapted and converged to form a new framework that also reflects two other elements which give further insight into agents, behaviours and vehicles concerning disinformation:

- The original *instigators* of disinformation, who may be different to the agents. These are the actors who initiate the creation and distribution of this content, often harnessing and paying for operationalisation. They are the real source and beneficiary of much disinformation. In some cases, the instigators can be the same as the actual implementing agents, but in many large-scale cases the latter may be paid or voluntary supporters or contractors, as well as unwitting participants. However, the functions of instigation and agency are distinct.
- The *intermediaries* that are vehicles for the message (e.g. social media sites and apps) - which allows for attention to the key role that they play in the dissemination and combating of disinformation, and how their systems may enable - or disable - implicated content, actors and behaviours.

This aggregation can be described with reference to **1. Instigators 2. Agents 3. Messages 4. Intermediaries 5. Targets/Interpreters** - creating the acronym **IAMIT**. This approach aims to capture the complete lifecycle - from instigation and creation to the means of propagation to real-life impact, through answering the following questions:

1. **Instigators:**

    - Who are the direct and indirect instigators and beneficiaries? What is their relationship to the agent(s) (below)? Why is the disinformation being spread? What is the motivation - e.g. political, financial, status boosting, misguided altruism, ideological, etc.? This includes, where discernible, if there is intent to harm and intent to mislead.

2. **Agents:**

    - Who is operationalising the creation and spreading of disinformation? This question raises issues of actor attribution (related to authentic identity), type ('influencer', individual, official, group, company, institution), level of organisation and resourcing, level of automation. Thus behaviours are implicated - such as using techniques like bots, sock puppet networks and false identities.

3. **Message**:

    - What is being spread? Examples include false claims or narratives, decontextualised or fraudulently altered images and videos, deep fakes, etc. Are the responses covering categories which implicate disinformation (eg. political/electoral content)? What constitutes potentially harmful, harmful and imminently harmful messaging? How is false or misleading content mixed with other kinds of content - like truthful content, hateful content, entertainment and opinion? How is the realm of unknowns being exploited by disinformation tactics? Are messages seeking to divert from, and/or discredit, truthful content and actors engaged in seeking truth (e.g. journalists and scientists)?

4. **Intermediaries**:

    - Which sites/online services and news media is the disinformation spreading on? To what extent is it jumping across intermediaries, for example starting on the 'dark web' and ending up registering in mainstream media?

    - How is it spreading? What algorithmic and policy features of the intermediary site/app/network and its business model are being exploited? Do responses seek to address algorithmic bias that can favour disinformation? Do the responses recognise that "...free *speech* does not mean free *reach*" because "there is no right to algorithmic amplification" and content moderation which may include limiting amplification should not be equated with the demise of freedom of expression online (DiResta 2018)? Also, is there evidence of coordinated behaviour (including inauthentic behaviour) exploiting vulnerabilities, in order to make it appear that specific content is popular (even viral) when in fact it may have earned this reach through deliberately gaming the algorithms?

    - Are intermediaries' acting in sufficiently accountable and transparent ways and implementing necessary and proportionate actions to limit the spread of disinformation?

5. **Targets/Interpreters:**

    - Who is affected? Are the targets individuals; journalists and scientists; systems (e.g. electoral processes, public health, international norms); communities; institutions (like research centres); or organisations (including news media);

    - What is their online response and/or real-life action? This question covers responses such as inaction, sharing as de facto endorsement, liking, or sharing to debunk disinformation. Is there uncritical news reporting (which then risks converting the role of a complicit journalist/news organisation from target into a disinformation agent).

    - Are there real-life impacts through actions? For example, such as influencing votes, promoting protests, inciting hate crimes, attacking journalists, and providing dangerous or erroneous medical advice, raising the question of whether responses engage with the wider context or are limited to the realm of the online content at hand.

Using this hybrid '**IAMIT**' framework as a starting point for conceptualising disinformation, it is then possible to categorise responses to disinformation on this basis. In particular, we can distinguish:

- Responses aimed at the instigators and agents of disinformation campaigns (Chapters 5.1, 5.2, and 5.3).

- Responses aimed at identifying disinformation, i.e. verifying messages in terms of falsity, exposing the instigators and agents. (Chapters 4.1, 4.2)

- Responses aimed at curtailing the production and distribution of disinformation and related behaviours, implemented particularly by intermediaries (Chapters 6.1, 6.2, and 6.3).

- Responses aimed at supporting the targets/interpreters of disinformation campaigns (Chapters 7.1, 7.2, and 7.3).

1.3.1 Disinformation and propaganda

Disinformation, as unpacked above, has distinctions from, and overlaps with, the notion of propaganda. Intentionality is core to an understanding of propaganda, in that the latter implies an organised, orchestrated campaign. This is not always the case with disinformation as broadly conceptualised in this study.

At the same time, as noted in the OED definition above, and in the Joint Declaration on Freedom of Expression and 'fake news', disinformation, and propaganda[14], disinformation may overlap with propaganda, which:

[14] https://www.osce.org/fom/302796

> *...is neutrally defined as a systematic form of purposeful persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for ideological, political or commercial purposes through the controlled transmission of one-sided messages (which may or may not be factual) via mass and direct media channels. (Nelson, 1996: p232-233)*

There is a long history where propaganda and disinformation are intertwined (Posetti & Matthews, 2018). Techniques of deceitful or 'dark' propaganda (e.g., selective use of facts, unfair persuasion, appeal to fear) are employed widely, e.g. in anti-EU campaigns, post-truth politics (Keane, 2018), ideology-driven websites (e.g., misogynistic or Islamophobic), and hyperpartisan media. This is often with the intent to effect actual behavioural changes e.g. to deepen social division, increase polarisation, influence public opinion, or shape key political outcomes.

While propaganda typically puts the emphasis on strategic persuasion towards action through harnessing narrative, identity and emotionality, the intellectual 'work' of disinformation (as conceptualised here) is to 'mess' with facts and knowledge in the primary instance (rather than target attitudes or behaviours). Propaganda typically harnesses disinformation to reinforce its bigger purpose. Yet, while disinformation can make a significant contribution to a propaganda package, these are not only analytically distinctive interventions, each can also stand alone. What complicates assessment is when disinformation is fused with propaganda techniques around linguistic, cultural, and national differences, such as to create new social barriers and amplify divisions. This fusion technique is a notable feature of divisive political campaigns, whether conducted internally within a State (e.g. campaigns with nationalistic objectives), or by foreign actors (e.g. designed to destablise political life in another State).

The rationale behind combining propaganda techniques with disinformation campaigns is to *enhance the credibility* of the message. It must be emphasised that the credibility of a message is separate from its veracity, since the former is about subjective perception of whether specific information seems credible, whereas verification is about evidence-based, independent assessment.

In addition, the merging of propaganda techniques and disinformation can be a strategy to move away from the use of patently false content in favour of using decontextualised, manipulative, and misleading content in order to distort the information ecosystem.

1.3.2 Disinformation and hate speech

Hate speech relies on group 'othering', and may engage disinformation as part of its arsenal, such as by reinforcing its message with false information and generalisations about a particular class of persons. Such speech may be part of a propaganda initiative, although not necessarily.

An important distinction needs to be made between disinformation on one hand and hate speech on the other, where hate speech is understood as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (UN, 2019; see also UNESCO, 2016). The two phenomena often intersect for instance when online abuse and disinformation are used hand-in-hand, such as in political smear campaigns, or misogynistic attacks on women journalists. They are nevertheless conceptually distinct, since false information can stand by itself and not touch on hatred, e.g. in anti-vaccination disinformation. Hatred, for its part, does not necessarily always implicate disinformation: it can rely simply on expressions of opinion and incitement without falsehoods, such as by amplifying fears, xenophobia, misogyny or other prejudices.

The focus of this study, in particular, is on the range of responses to disinformation assessed through a freedom of expression lens. Therefore, responses purely focused on hate speech are out of scope. Where responses to disinformation are tied up with hate speech, however, the phenomenon is examined from that perspective.

1.4 Disinformation, freedom of expression, and the UN sustainable development goals

Seventeen global Sustainable Development Goals (SDGs) were set by the United Nations General Assembly in 2015[15]. A number of them are impacted by use of broadband technology and internet communications company services for the proliferation of online disinformation - but also for the implementation of some categories of disinformation responses. These are:

- **SDG 16 on peaceful and inclusive societies and SDG 5 on gender equality**:
  - **Online disinformation is** often used to target individuals (such as politicians, journalists, human rights defenders), governments, groups such as ethnic minorities, women and gender identity-based communities, and religious congregations and identities, including in messages which may lead to violence, hatred, and discrimination.
  - The **algorithms used by social media and search engines** to prioritise and recommend content (including disinformation) have been shown to prioritise and recommend content that is attention- and engagement- focused, and prone to bias, accordingly potentially working against inclusivity. (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018a)
- Of particular relevance to this report on Freedom of Expression (FoE) and disinformation is **SDG 16.10 on public access to information and fundamental freedoms**
  - Citizens' rights to express themselves freely and participate on an informed basis in online societal debates are jeopardised by online disinformation, especially when distributed at scale. False content can undermine citizens' beliefs and trust in facts, science and rationality, and therefore stoke cynicism about online information that contradicts their opinions. This can deter public participation, and impact negatively on the exercise of rights and obligations concerning civic actions. This is especially relevant for citizens and communities targeted in disinformation campaigns using hate speech as a tool to fuel division and inflame tensions.
  - It is noteworthy that politicians and governments are among the main instigators and vectors of disinformation (Brennen et al 2020; Bradshaw & Howard 2019).

[15] https://sustainabledevelopment.un.org/?menu=1300

- The rising use of AI algorithms for automatic content filtering of disinformation (and other kinds of content) can lead to over-censorship of legitimate content, thus infringing on the author's freedom of expression and right to access information. These algorithms can also exhibit inherent biases and be prone to manipulation.

- Orchestrated and organic disinformation campaigns targeting journalists (particularly women journalists) and news outlets as a means of undermining citizens' trust in journalists and journalism as credible and independent sources of information.

- Another example is disproportionate legal responses to disinformation which can sometimes lead to internet shutdowns and censorship, inhibiting reporting, and criminalising journalism, as well as vague legal definitions of disinformation which can be used to silence political opposition or dissenting voices (e.g. via 'fake news' laws)

- **SDG 4 on inclusive and equitable quality education:**

  - As citizens are increasingly using the internet and search engines to find information for educational purposes, **high levels of online disinformation** can seriously impact on the knowledge processes essential for the quality of education, as many learners are unable to judge the trustworthiness of online sources and the veracity of the information they find online. This has become increasingly important as COVID-19 has forced so much education online.

  - The search engine algorithms used by citizens to find information can be gamed to prioritise viral disinformation, which in turn can lead to learners (especially children and older generations) starting to believe in conspiracy theories and other false or low-quality online information.

  - On the positive side are investigative journalism projects focused on disinformation and **media and Information literacy initiatives**, **including data literacy**, designed in response to online disinformation that aim to impact positively on citizen education, knowledge, and abilities to identify and protect themselves from disinformation.

- **SDG 3 on healthy lives and promotion of well-being for all ages**:

  - Health-related disinformation in general - as demonstrated during the COVID-19 pandemic and including long-standing anti-vaccine propaganda - jeopardises citizens' healthy lives and well-being (e.g. diet-related disinformation). As a result of anti-vaccine disinformation, vaccine take-up rates have shown a sharp decline in recent years (e.g., in Africa (France 24, 2020), Asia (Power, 2020), Europe (Larson, 2018) and North America (Burki, 2019)).

Disinformation runs counter to the agreed SDGs. Yet, its purveyors (wittingly or unwittingly) and operating with a range of motives, still foresee advantage in investing time and resources in producing and circulating it - leveraging the business models and technologies of internet communications companies and the news media to do so.

> *"At the same time, disinformation is a 'game' with no long-term winners. Escalating the volume of disinformation in play ultimately devalues facts for everyone and puts humanity on a path towards ubiquitous ignorance. The achievements of civilisation based upon freedom of expression to date are being jeopardised as a result. At stake are issues of health, democracy, financial security, the environment, peaceful resolution of social conflict, social cohesion, and more."*

Disinformation is a phenomenon that is too challenging for any single state or company to manage in isolation - it requires collaboration with researchers, civil society and the news media. Paid microtargeting of individuals with disinformation content is one example that calls out for unprecedented cooperation; the peer-to-peer use of closed social messaging networks that spread falsehoods is another.

It is for this reason that this study examines the range of responses that can prevent, inhibit and counter disinformation. The following chapters assess the full suite of possibilities, and their respective strengths and weaknesses, as well as potential risks to freedom of expression rights, as multiple actors seek to tackle disinformation.

The next chapter - Chapter two - introduces the typology of disinformation responses which forms the backbone of this study. Chapter three provides a detailed mapping of existing research, highlighting knowledge gaps and opportunities for further study. Then, each of the eleven response types presented in the original taxonomy devised for this study is systematically analysed.

1.5 Methodology

The findings presented here are the result of desk research carried out (September 2019-July 2020) by a multidisciplinary team of international authors who worked in a highly collaborative fashion.

The research for this study sought to include sources pertaining to countries on all continents, including where possible (according to the language capabilities of the researchers), materials in languages other than English. The libraries and databases of academic institutions, individual States, civil society organisations and news media websites were targeted by the researchers. Many of these collected sources have now been aggregated into the study's bibliography.

An Expert Oversight Group comprised of Associate Professor Fabrício Benevenuto, Federal University of Minas Gerais; Prof Divina Frau-Meigs, Université Sorbonne Nouvelle - Paris 3; Prof Cherian George, Hong Kong Baptist University; Dr Claire Wardle, Executive Chair of First Draft; and Prof Herman Wasserman, University of Cape Town provided feedback. The research team also worked closely with the UNESCO secretariat to shape this study.

2 Typology of Disinformation Responses

Authors: Kalina Bontcheva, Julie Posetti, Denis Teyssou, Trisha Meyer, Sam Gregory, Clara Hanot, Diana Maynard

This chapter introduces the hierarchical typology of disinformation responses elaborated as part of the research carried out for this report.

According to this taxonomy, disinformation responses are categorised by their aim of targeting particular aspects of the problem, rather than in terms of the actors behind them (e.g. internet communication companies, governments, civil society, etc.). Framing enables identification of the complete set of actors involved in, and across, each category of disinformation response. For example, even though at present many actors tend to act independently and sometimes unilaterally, such a response-based categorisation can point out possible future synergies towards a multi-stakeholder approach to delivery within and across categories of intervention.

A second key motivation behind this response-based categorisation is that it allows for an analysis of the impact of each response type on freedom of expression (and, where appropriate, on other fundamental human rights such as privacy). In particular, each response category is evaluated not only in terms of its general strengths and weaknesses, but specifically in relation to freedom of expression.

The typology of disinformation responses distinguishes four top-level categories (see Figure 1 below):

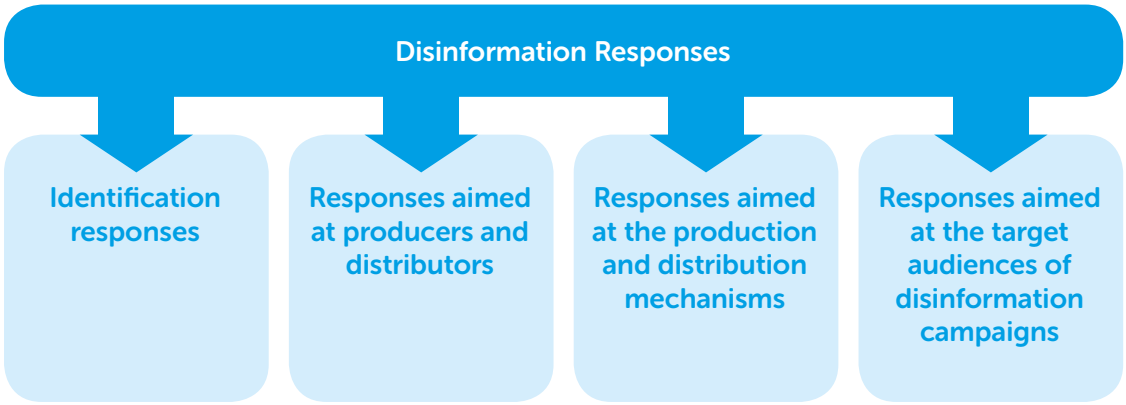


Figure 1. *Top-level categories of disinformation responses*

The categories in this typology are not always mutually exclusive. That is, there are some interventions that belong to more than one response category typology, even if there are dimensions that encompass other categories, for example. Where this is the case, they are addressed under one of the categories but cross referenced in other chapters where relevant. For example, election-specific fact-checking initiatives are relevant to the chapter discussing electoral-oriented responses (5.3) and the chapter on fact-checking responses (4.1), so they are dealt with primarily in chapter 5.3, but also referenced in chapter 4.1.

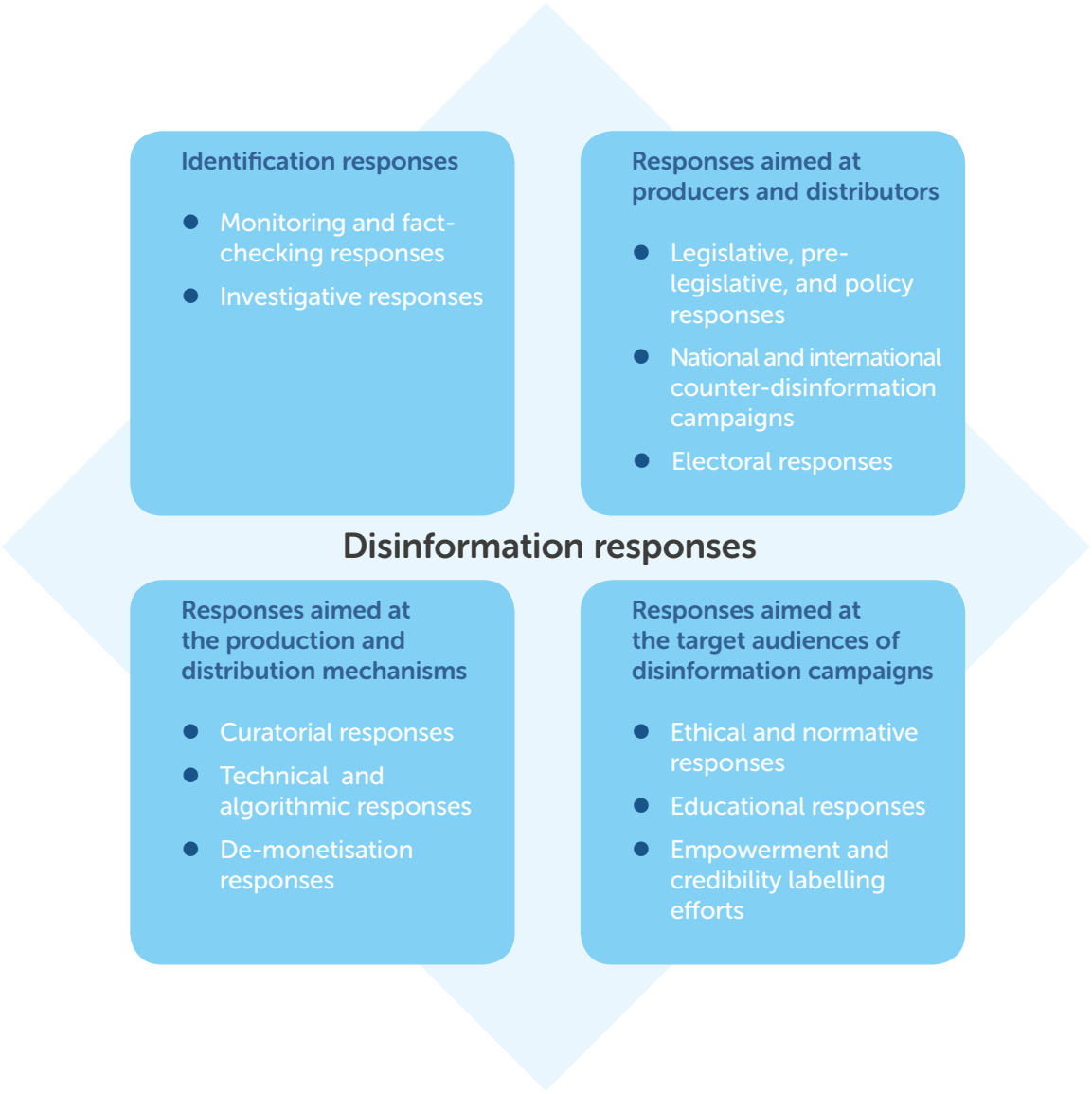


Figure 2. *The 4 top-level response categories and their eleven sub-categories.*

In more detail, **identification responses** involve monitoring and analysis of information channels (e.g. social media and messaging, news media, websites) for the presence of disinformation. The objective here is to pinpoint the existence and extent of disinformation. In particular, two subtypes of identification responses are recognised:

- **Monitoring and fact-checking responses**, which tend to be carried out by news organisations, internet communications companies, academia, civil society organisations, and independent fact-checking organisations, as well as (where these exist) partnerships between several such organisations.

- **Investigative responses**, which go beyond the question of whether a given message/content is (partially) false, to provide insights into disinformation campaigns, including the originating actors, degree of spread, and affected communities.

The second umbrella category captures **responses aimed at producers and distributors of disinformation through altering the environment that governs and shapes their behaviour** (law and policy responses):

- **Legislative**, **pre-legislative**, **and policy responses**, which encompass regulatory interventions to tackle disinformation.
- **National and international counter-disinformation campaigns**, which tend to focus on the construction of counter-narratives.
- **Electoral responses** are designed specifically to detect, track, and counter disinformation that is spread during elections. Even though there are other important targets of online disinformation (e.g. vaccination and other health disinformation), a separate category is introduced for responses specific to countering election disinformation due to its impact on democratic processes and citizen rights. This category of responses, due to its very nature, typically involves a combination of monitoring and fact-checking, legal, curatorial, technical, and other responses, which will be cross-referenced as appropriate. This highlights the multi-dimensional approach required in order to combat election-related disinformation, with its specific potential to damage the institutions of democracy.

The third broad category brings together **responses within the processes of production and distribution** of disinformation, which include curation, demonetisation, contextualisation and use of automation:

- **Curatorial responses** address primarily editorial and content policy and 'community standards', although some can also have a technological dimension, which will be cross-referenced accordingly.
- **Technical and algorithmic responses** use algorithms and/or Artificial Intelligence (AI) in order to detect and limit the spread of disinformation, or provide context or additional information on individual items and posts. These can be implemented by the social platforms, video-sharing and search engines themselves, but can also be third party tools (e.g. browser plug-ins) or experimental methods from academic research.
- **De-monetisation responses** are designed to stop monetisation and profit from disinformation and thus disincentivise the creation of clickbait, counterfeit news sites, and other kinds of for-profit disinformation.

The fourth umbrella category clusters **responses aimed at supporting the target audiences of disinformation campaigns** (i.e. the potential 'victims' of disinformation). Such responses include guidelines, recommendations, resolutions, media and data literacy, and content credibility labelling initiatives. These different responses are sub-classified into:

- **Ethical and normative responses** carried out on international, regional and local levels involving public condemnation of acts of disinformation or recommendations and resolutions aimed at thwarting these acts and sensitising the public to the issues.
- **Educational responses** which aim at promoting citizens' media and information literacy, critical thinking and verification in the context of online information consumption, as well as journalist training.

- **Empowerment and credibility labelling efforts** around building content verification tools and web content indicators, which are practical aids that can empower citizens and journalists to avoid falling prey to online disinformation. These efforts may also be intended to influence curation in terms of prominence and amplification of certain content – these are included under curatorial responses above.

After a detailed literature review and landscape mapping exercise in chapter three, this report turns to defining, analysing and evaluating disinformation responses according to this categorisation. In each case, the idiosyncratic properties of the category are detailed and a common set of questions isasked to trigger explication of the underpinnings of each response type. These questions are:

- Who and/or what does the response type monitor?
- What is the target audience of the response type/whom does it try to help?
- What are the outputs of this response type (e.g. publications, laws)?
- Who are the actors behind these responses, and who funds them (where known)?
- How is the efficacy of these responses evaluated?
- What is their theory of change?
- What are their strengths and weaknesses in general, and with respect to freedom of expression in particular?
- What are the gaps and potential synergies identified in the course of the analysis?

Finally, where relevant, the COVID-19 'disinfodemic' (Posetti and Bontcheva 2020a; Posetti and Bontcheva 2020b) is addressed through a mini case study within the chapters.

Research Context and Gaps

3

Authors: Diana Maynard, Julie Posetti, Kalina Bontcheva and Denis Teyssou

This chapter situates the disinformation focus of this report within the context of existing theoretical frameworks and prior reports on this topic. It also relates disinformation to freedom of expression and the Sustainable Development Goals relevant to the Broadband Commission (specifically SDG 16 on peaceful and inclusive societies, and SDG 16.10 on public access to information and fundamental freedoms). In particular, the focus is not only on the false content itself, but also the actors, their motivations for sharing disinformation, and the targets of disinformation campaigns, thereby including discussion of the amplification and manipulation of this kind of content. Additionally, the chapter examines the literature regarding modalities of response to disinformation. Then, it discusses in more depth the gaps in the research carried out prior to early 2020, especially in relation to defining the novel contributions of this study, compared with previous reports on the manifestations of disinformation.

With this gap analysis, special attention is paid to the impact of disinformation on societies and its reception by the public, by reviewing literature in cognitive science and social psychology, in addition to that found in the fields of politics, journalism, information and communication sciences, and law. The review encompasses not only academic literature, but also policy reports from industry and civil society groups, white papers, books aimed at the mainstream public, and online news and magazine articles. It should be emphasised, however, that this review is not intended to be exhaustive, rather it is designed to map some of the key research trends and gaps, while also identifying gaps in responses.

The chapter does not attempt to definitively assess the quality of the selected works, but rather to understand the nature of a range of existing research on disinformation, the theoretical underpinnings, types of studies that have been carried out, and the ways in which disinformation has been discussed in journalistic, academic and official (i.e. State and non-State actors) circles, as well as how this has been disseminated more widely to the general public. It then summarises the findings, focusing on some key areas such as political disinformation and policymaking, and highlights some emerging trends, before discussing the limitations of such research and situating this report within the scholarship.

The aim of this review is thus to understand what can be learnt about what kinds of disinformation exist; who instigates and disseminates it and why; at whom the information is targeted; the relative effectiveness of different kinds of disinformation and different modalities; existing responses to it; and what recommendations have been made for the future, in particular in the light of freedom of expression concerns. This paves the way for the following chapters which investigate in more depth the various responses to disinformation, as well as the theories of change associated with them, and possible benefits and drawbacks.

Attention is given to highlighting new threats, such as undermining freedom of expression by indiscriminately using Artificial Intelligence (AI) filtering methods, and to the rise of synthetic media (also called 'deepfakes') as new modes of disinformation. The latter problem can already be seen in practice, where several politicians and journalists have been targeted and smeared with inappropriate sexual misconduct allegations in manipulated and/or deepfake videos.[16]

A related recent trend, which has been largely underestimated in the past, is the rise of adversarial narratives (Decker, 2019), whereby disinformation strategies include not only simple conspiracy theories and outright lies, but also involve more complex phenomena

[16] See examples from Finland (Aro, 2016), India (Ayyub, 2015), and South Africa (Haffajee & Davies, 2017), among others.

whereby true and false information is emotionally charged and deliberately entangled in intricate webs designed specifically to confuse, shock, divert and disorientate people, keeping truth-seekers always on the defensive. If information is a condition for public empowerment, then disinformation can be seen to function in terms of displacing and discrediting information, often with the rationale of disempowerment and driving confusion. One example of this is 'gaslighting', a powerful strategy aimed at control through power and manipulation of people's perceptions of reality - thereby generating fears and sowing disruption, and then appearing to offer solutions (Keane, 2018). These disinformation techniques, often described as the "weaponisation of information", can destroy social cohesion and threaten democracy (Hansen, 2017). They can stimulate public demand for stronger certainty and greater political control, thereby risking further curbs on freedom of expression, and strengthening social authoritarianism (Flore et al., 2019). On the other hand, responses to disinformation are developing in sophistication and incorporating human rights standards in order to counter the potential harms at stake.

3.1 Conceptual frameworks for understanding contemporary disinformation

In recent years, there has been a flurry of research investigating not only the nature and extent of disinformation, but also the psychological underpinnings and theoretical frameworks. These frameworks capture different aspects. An overarching view is taken by Wardle and Derakhshan (2017), who consider 'information disorder' as a tripartite problem where (in their definitions) 'disinformation' sits alongside 'misinformation' and 'mal-information'. Other views range through to narrower classification systems such as the political disinformation campaign characterisation of the Digital Forensic Research Lab (Brooking et al., 2020).

3.1.1 "Information disorder" and "information warfare"

In their report for the Council of Europe, Wardle and Derakhshan (2017) elaborate their concepts and provide a background summary of related research, reports and practical initiatives produced around this topic up to the middle of 2017. Their report investigates ideas and solutions for, and from, the news media and social media platforms, as well as examining future directions and implications. This includes focus on the use of AI, not only for detecting disinformation but also for creating it. The report also details 34 recommendations for technology companies, governments, media organisations, funding bodies, and broad citizenry. Many of these recommendations are already in place in some form (for example, some technology companies are already building fact-checking tools. Some recommendations lend themselves to further unpacking (for example, how civil society could "act as honest brokers", or how education ministries could "work with libraries").

Wardle and Derakhshan's conceptual framework follows on from their previous work "Fake News, It's Complicated" (Wardle, 2017a), which defines seven types of mis- and dis-information, ranging from satire and parody (which, being mis-interpretable, have the potential to lead to what they call mis-information) through to full-blown deliberate fabrication. The framework situates the production and distribution of disinformation

as a tripartite process consisting of Agent, Message, and Interpreter (target). However, as signalled in the Introduction to this study, the practicality of this frame encounters the challenge of distinctions between mis-information and dis-information being based primarily on motive and awareness of falsity. Motives are not only diverse and often contradictory, but also frequently not clear. Furthermore, the distinction may risk over-emphasising intentionality at the expense of commonality of effect. For example, if people decide against vaccination through engagement with false content, the consequence is the same, whether the mode of transmission is mis-information or dis-information. Where motives become significant as an issue, although they are often hard to pinpoint, is in assessing the appropriateness of a given response with respect to how it establishes the issue of motives at hand. That is why this study pays attention to investigative responses as a source of knowledge for informing other types of responses.

Another consideration related to the 'information disorder' framework is that it can favour a binary distinction between information that is 'ordered' or 'disordered', and thereby reinforce a view that reduces the veracity of communications to issues of black and white, while overlooking or denying the range of unknowns, whether these be scientific uncertainties or legitimate policy debates. Another issue is that 'mal-information' could be interpreted in a way that stigmatises a range of non-disinformational narratives, which intrinsically select and interpret particular facts and contexts as part of their legitimate contestation around meaning.

In this light, the research in this study operates at a more abstract level than privileging categories of false or misleading content through the criteria of motives, and instead puts the focus on all false content that has the potentiality of defined harm. This provides a means towards assessing the full range of responses as they conceptualise themselves.

A strategically focused approach to the issue of disinformation is assessed by Derakhshan (2019) in his report "Disinfo Wars". This discusses the relationship between agents and targets in what he calls a "taxonomy of information warfare". Accordingly, the approach directs the idea of disinformation into a much narrower concept that articulates to political and even military strategy. An example of the latter is the perspective on 'Information Operations' / 'Influence Operations' taken by the Rand corporation, which links these terms to "the collection of tactical information about an adversary as well as the dissemination of propaganda in pursuit of a competitive advantage over an opponent".[17] A similar position is adopted by the European External Access Service (EEAS) East Stratcom Task Force[18]. Derakhshan argues that the majority of money and effort spent on countering disinformation in "information warfare" should be focused on those who are targeted, i.e. non-state actors like the media.

While his argument covers a wide range of activities, it focuses to some extent on false content distributed with a particular motive, as with Wardle's earlier work (Wardle, 2017a). As discussed above, this is complicated operationally, and it goes beyond even the complex issues of attribution. In addition, while strategic focus on geopolitical dimensions and particularities is important, society also faces the issue of disinformation as a far wider problem. There is also a lack of evidence that work with one constituency (the media, or the general public) is less or more effective than work with another.

17 https://www.rand.org/topics/information-operations.html
18 https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1006

3.1.2 Political disinformation campaigns

A perspective that relies less on warfare metaphors but deals with political disinformation as a broader concept has been adopted by researchers at the Digital Forensic Research Lab and Google's Jigsaw (a division that includes a focus on combatting the 'unintended consequences' of digital technology) has proposed and tested a classification system for political disinformation campaigns, built on 150 variable options (Brooking et al., 2020). The main aim of this framework is to enable the description and comparison of very different kinds of political disinformation efforts. The scheme has six major categories: target, platform, content, method, attribution, and intent. Each of these is broken down into further categories and subcategories. The table below shows the first and second level categories, with some examples of the third level. Typically, the third level categories are binary (e.g. whether it is a government-related target or not), although the quantitative measures involve numbers or ratios, and some have free-form responses. In addition, all second level categories have a category where free-form notes can be added, and some also have an "other" subcategory.

1st level	2nd level	Notes / Examples
Target	Primary target	Government, political party, business, racial group, influential individuals (including journalists) and groups of individuals, etc.
	Quantitative measures	Indicators/rankings of political stability, internet freedom, refugee counts etc.
	Concurrent events	War, elections etc.
	Secondary target	Rarely used
	Tertiary target	Rarely used
Platforms	Open web	State media, independent media, other
	Social media	Facebook, Instagram, Twitter, forums, etc.
	Messaging platforms	WhatsApp, Telegram, Wechat, SMS, etc.
	Advertisements	(Purchased by disinformants to disseminate a message of disinformation, including on social media and the open web)
	Email	
Content	Language	
	Topics	What the disinformation is about, e.g. government, military, elections, terrorism, racial, etc.
Methods	Tactics	Brigading, sock puppets, botnets, search engine manipulation, hacking, deepfakes, etc.
	Narrative techniques	Constructive (e.g. bandwagon, astroturfing); Destructive (e.g. intimidation, libel); Oblique (trolling, flooding)
Attribution	Primary Disinformant	Country, bloc, other
	Disinformant Category	As for target category, e.g. government, political party, business, influential individual, minority group
	Quantitative Measures	As for target, e.g. political stability data, internet freedom, refugee counts
	Concurrent Events	As for targets, e.g. war, elections, etc.
	Secondary Disinformant	Rarely used
	Tertiary Disinformant	Rarely used

Intent	Object	Free text (1 or 2 short sentences)
	Category	e.g. civil, social, economic military

Table 1. *Simplified version of the political disinformation campaign categorisation scheme devised by Brooking et al. (2020).*

In this work, Brooking et al. define political disinformation as "disinformation with a political or politically adjacent end", which captures "disinformation spread in the course of an election, protest, or military operation, as well as "the widespread phenomenon of political 'clickbait' disseminated for personal financial gain".

Their framework defines a political disinformation campaign as "political disinformation that demonstrates both coordination and a discrete objective." They note that, first, objectives may not always be obvious, even though they must exist; and second, that campaigns with changing objectives can thus become discernibly distinct from each other (i.e. if the objective changes, it becomes a new campaign). Furthermore, they note that political disinformation campaigns almost always involve what they call "amplification of content". This concept, which is discussed in more detail in the following section, is termed "political astroturfing" by Keller et al. (2019), "coordinated inauthentic behavior" by Facebook (Gleicher, 2018a), and noted as a feature of 'astroturfing' in the targeting of journalists with misleading information designed to "mislead, misinform, befuddle, or endanger journalists" by Posetti (2013). Not all instances of this constitute disinformation as such, but there is a clear overlap since the aim is to create an "illusion of consensus or popularity," and in some instances, to inflict harm. Some researchers have tried to capture this complex interplay through a "matrix of disinformation harms", which encompasses polarisation and radicalisation along one dimension and propaganda and advertising along the other (Frau-Meigs, in press).

In providing a basis for comparing different kinds of disinformation, this framework also has the benefit of enabling detailed background information to be represented. Understanding the situational context such as the presence of military conflict, or levels of political stability may help with both short and long-term assessment and the provision of appropriate solutions. However, it also risks the case that some of the factors may be unknown or irrelevant. As with other frameworks discussed, notions of intentionality and attribution are also not always evident. As significant as deliberate disinformation is during such political campaigns, this study bears in mind the wider picture that includes unintentional falsehoods in play (such as health issues), and therefore maintains a focus that covers responses wider than those dealing with political issues.

3.1.3 Information influence

Similar to the political disinformation campaign characterisation, the Handbook for Communicators (Pamment et al., 2018) views disinformation in the context of the wider sphere of "influence activities" and from the point of view of policymaking (in the case of that handbook, the Swedish government). This framework deconstructs influence activities conducted by foreign powers, focusing on rhetorical strategies, techniques, and influence stratagems, and aims to enable policymakers to identify, understand, and counter these increasingly sophisticated activities and campaigns. This approach focuses particularly on safeguarding society's "democratic dialogue", which they explain as "the right to open debate, the right to arrive at one's own opinion freely, and the right to free expression". In this light, they view methods of social resilience, such as informing and educating the public, as the foundation for combatting disinformation and influence

activities, and they focus their attention on public communicators within governments and state organisations accordingly.

'Information influence' in this framework is closely related to disinformation, which Pamment et al. define as "a technique based on the distribution of false information intended to mislead and deceive". The authors argue that those who conduct "influence activities" are only a step away from (perfectly legitimate) advertising campaigns which attempt to sway people to buy a product, for example. They argue that it is precisely the notion of openness that differentiates them: advertising and public relations should be transparent in their motives, and follow clear rules; on the other hand, information influence involves the covert and deceptive deployment of false content. In this regard, the approach of Pamment et al. overlaps substantially with broader uses of the term 'information operations' such as as references to the combination of co-ordinated and inauthentic behaviour (such fake profiles and hidden behaviours) as a wider phenomenon than cases of military or geopolitical deployments.

> *"Given that societies are built on trust, deceptive 'information influence' undermines the democratic principle of the public's right to know and access information"*

Given that societies are built on trust, deceptive 'information influence' undermines the democratic principle of the public's right to know and access information, potentially destabilising democracy by muddying the informational waters so much that it becomes impossible to discern accurate information from falsehoods, and credible journalism from propaganda, broadly undermining trust in public interest information. In this regard, the concept of 'information influence' also resonates in part with the concept of infodemic[19] popularised by the World Health Organisation, and which designates "an overabundance of information – some accurate and some not – occurring during an epidemic. It makes it hard for people to find trustworthy sources and reliable guidance when they need it."[20]

The theory of information influence adopted by Pamment et al. has three parts: awareness, identification, and countering.

Awareness consists of understanding the anatomy of an information campaign, as well as the process of opinion formation. In this light, information influence can be distinguished by three main features: it is deceptive, intentional, and disruptive. It should be noted, however, that these aspects are not always easy - or even possible - to determine, signalling an important gap in this theory. As previously discussed, intentionality can be hard to determine, or at least to attribute, and the extent and impact of disruption is hard to measure.

The process of **identification** of 'information influence' is based on the idea of strategic narratives, which can be seen as a deliberate manipulation of some fundamental belief such as that the earth is round.[21] Distinct from other frameworks such as those of Derakhshan (2019) and Brooking et al. (2020), target groups here are always the public,

[19] https://www.merriam-webster.com/words-at-play/words-were-watching-infodemic-meaning; https://www.who.int/teams/risk-communication/infodemic-management

[20] https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference

[21] Note that since disinformation, as conceptualised in the approach of this study, could count as one of a number of different types of information influence, this does not signify that all strategic narratives equate to disinformation, nor that all strategic narratives are fundamentally deceptive.

and can be broken down into general public, socioethnic groups (e.g. a religious group), and psychographic groups (those with specific personality traits).

In the framework of Pamment et al., disinformation is defined far more narrowly than it is treated in this report. It is classified as a technique distinct from techniques involving technical exploitation, which includes bots, 'deepfakes', and sock puppet networks. These in turn are seen as distinct from the category of "deceptive identities", which includes what they term "fake media"[22] and the loosely defined "counterfeits". The other three categories - social and cognitive hacking, malicious rhetoric, and symbolic actions, are more loosely related to disinformation, encompassing notions of bias such as filter bubbles, strawman tactics, and leaking and hacking, respectively. On the other hand, satire and parody are (problematically) classified as disinformation. In contrast, in this study, it is recognised that while disinformation is often orchestrated, it is not per se a technique – instead, it makes use of techniques like technology and deceptive identity. The same point applies to the analysis of Francois (2019), which comes close to elevating behaviours, including inauthentic behaviours (and fake actors), to being defining features of what should be considered as disinformation. While such trademark signs of disinformation are significant, this study also recognises that many cases of disinformation also exist without these features.

The framework by Pamment et al. faces the challenge, like many already discussed, of the practical ability to make distinctions given reliance on assumptions about motive and intent. This challenge also applies to those who interpret behaviours as a barometer of motives, in that there are complex levels between, for instance, a person who shares false content believing it to be true and helpful, and an agent who amplifies it, and further compared with an instigator operating with a wider strategy. On the other hand, the Pamment et al. assessment does avoid a potential pitfall of the concept of 'mal-information', in recognising that not all persuasive or strategic narratives equate to disinformation.

Finally, in terms of strategies for **countering** information influence, Pamment et al. suggest four categories, ordered temporally. The first responses are the two fact-based techniques of assessment and then informing. These are followed by two advocacy-based techniques and, lastly, defence. The first step, assessing the situation, can involve methods such as fact-checking and investigating the transparency of the information. Informing involves steps such as making statements to signal issues, and correcting factual inaccuracies. Advocating is described as use of mechanisms such as dialogue and facilitation. Defence is the final stage in the process which involves official blocking, reporting, and removal of disinformation. While not approaching the extent of responses covered in this study, the Pamment et al. framework does have the merit of highlighting the links between awareness, identification and response.

[22] Editors' note: The terms 'fake news' and 'fake media' are problematic and should be avoided where possible because they are frequently weaponised as tools in the disinformation 'arsenal' in an attempt to discredit journalists and news reporting by actors seeking to chill accountability journalism. See UNESCO's *Journalism, 'Fake News' and Disinformation* for further discussion (Free to download here in multiple languages: https://en.unesco.org/fightfakenews)

3.2 Empirical and applied research

Moving on from theoretical frameworks which attempt to define and classify various kinds of disinformation and, in some cases, potential responses to it, this chapter now focuses on more empirical and applied research, looking at some key trends and examples of specific case studies.

Bradshaw and Howard's "Global Inventory of Organised Social Media Manipulation" (2019) focuses on social media manipulation by governments and political parties. Their report analyses the trends of what they call 'computational propaganda', looking at tools, capacities, strategies, and resources. Their surveys show that in recent years, evidence of organised social media manipulation campaigns is becoming more widespread worldwide, with the number of countries involved increasing by 150% in two years. In 2019 they found evidence of such campaigns in 70 countries, up from 48 countries in 2018 and 28 countries in 2017, with Facebook being the most common social media source.

Martin and Shapiro (2019) also present a detailed classification system for online "foreign influence" initiatives, which compares the characteristics, tactics, rhetoric and platform choices of different attackers. A few studies have attempted to dig deeper into the underlying motives of these kinds of initiatives, but these are restricted to country-specific case studies. Ong and Cabañes (2018) investigate, from an "ethnologically informed" perspective, the motivations and behaviour of those who are recruited to produce networked disinformation and social media manipulation in the Philippines, while Chaturvedi (2016) investigates similar issues in India.

However, despite these and other reports discussing these forms of organised political disinformation and 'influence operations', there remains a lack of coordinated in-depth research into this phenomenon as a whole, especially at more than a case- or country-specific level. These systems can influence people sufficiently to change their votes, buy products and change perceptions - sometimes with enormous consequences for democracy or public health. So-called 'dark PR' has been defined as the "manipulation at scale for money without any concerns for the damage to the planet, country, or even individual safety"[23], leading to a worldwide industry of PR and marketing organisations buying services that use fake accounts and false narratives to spread disinformation via end-to-end online manipulation systems (Silverman et al., 2020).

A number of countries around the world have sought to make it a crime to create and distribute disinformation of this type (Adhikari, 2020), although the definitions of what is acceptable vary substantially. In practice, finding the sources and proving intent may not be a trivial process for either law enforcement agencies or companies themselves. Adhikari notes that Facebook has attempted to curb such disinformation spreading practices, banning in 2019 a number of dark PR firms for attempting to influence elections, or for what it calls "coordinated inauthentic behavior" in various countries. However, these kinds of activity are still widespread, and new companies promoting such services can be easily set up.

[23] Definition by Rob Enderle, principal at the Enderle Group, quoted in the E-commerce Times article '*Black PR' Firms Line Their Pockets by Spreading Misinformation* by Richard Adhikari: https://www.ecommercetimes.com/story/86444.html

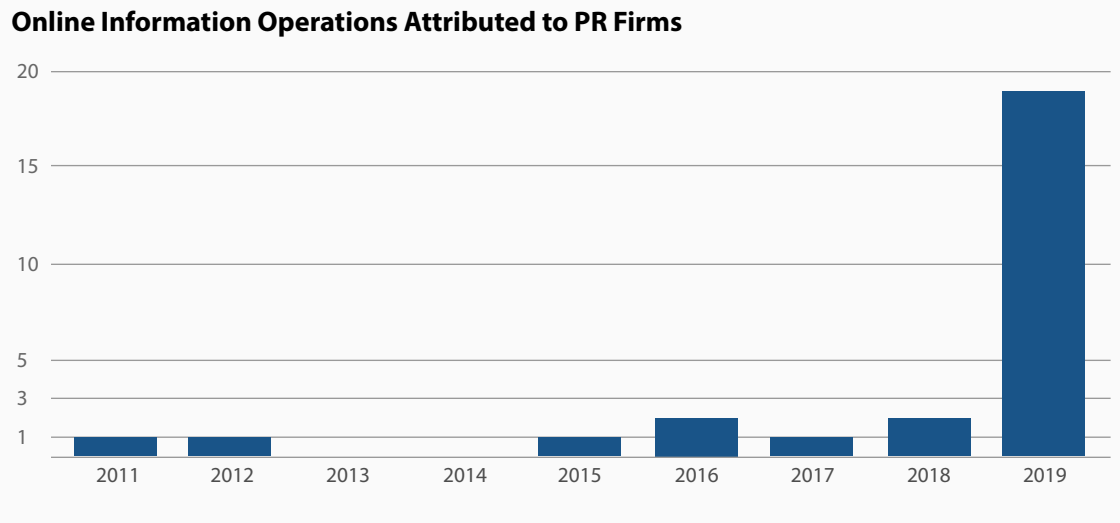


Figure 3. *Chart Source*: *Buzzfeed News* (*Silverman et al., 2020*)

From the BBC, analysis of misleading stories during the COVID-19 pandemic resulted in a typology of seven kinds of actors involved in generating disinformation (Spring, 2020). Other noteworthy journalistic investigations giving insight into the agents and instigators include those from Buzzfeed, Codastory and Rappler, for example (Dorroh 2020) - as discussed further in chapter 7.1. At the time of writing, however, there was a scarcity of detailed academic studies on this phenomenon, and methods for preventing it at its source were not obvious.

While notions of 'influence operations' are not themselves new, the proliferation of these in 2019, as illustrated below, requires urgent attention. On the other hand, it should be noted that not all 'influence operations' necessarily equate to the characterisation of disinformation used by this study, in the sense that some such initiatives may not harness false or misleading content, nor rely on inauthentic behaviour. Recent high profile cases concern mechanisms such as "inciting racial tension" (Neate, 2017) and "co-ordinated inauthentic behaviour" (Gleicher, 2019a) which leave open a number of possibilities as to their harnessing of disinformational content. Some coordinated campaigns can be mounted with accurate content, transparent behaviours and authenticated actors, as for example in advocacy by some civil society groups, public health communications by governments, and public relations initiatives by companies. The topic of organised influence therefore needs to be approached with appropriate nuance when researched from the point of view of when and how it intersects with false and misleading content with potential to harm.

3.2.1 Social and psychological underpinnings

A strand of research into disinformation situates it within its social and psychological context in order to define and understand appropriate responses. Even if some of the mechanisms of disinformation are new, responses to them can/may be guided by the decades of research into human cognition. As will be discussed elsewhere in this study, it can be hard to persuade people who want to believe a piece of information that this content is indeed false - or that a false 'fact' can make a difference to the meaning they attribute to a bigger picture. Even if fact checkers disprove false information, research has shown that it can be extremely difficult to change people's minds on misconceptions, especially if they believe there is even a kernel of truth within the falsity (Johnson & Seifert,

1994; Nyhan & Reifler, 2010). As the economist J.K. Galbraith once wrote: "Faced with a choice between changing one's mind and proving there is no need to do so, almost everyone gets busy with the proof" (Galbraith, 1971). Repetition and rhetoric are powerful devices: people are more likely to believe information when they see it repeated over and over again (Zacharia, 2019). Importantly, according to Effron & Raj (2019), such repeated exposure means that such people also have fewer ethical concerns about resharing it, whether they believe it or not.

Longstanding research in political science has shown the power of rhetoric time and again (Kroes, 2012; Grose and Husser, 2008): linguistically sophisticated campaign speeches by election candidates are far more likely to influence people to vote for them. This linguistic sophistication involves presenting the message - no matter what its content - in a tailored rhetorical way that also conveys emotional resonance. However, one finding has been that while linguistic sophistication (i.e. presenting the message in a particular rhetorical way, rather than changing the message itself) is more likely to persuade those with higher education, it does not dissuade those without (Grose and Husser, 2008). While campaign speeches as such should not be equated with disinformation, these findings lead to the observation that disinformation combined with non-informational dimensions (emotional quotients) could be more powerful than when it is presented alone.

Taking this a step further, others frame relevant aspects of disinformation within the notion of "psychological warfare" (Szunyogh, 1955, cited in Cordey, 2019). Guadagno and Guttieri (2019) provide an overview of research in psychology and political science in this context through the spreading of disinformation. They review a number of social, contextual and individual factors that contribute to its proliferation. Focusing specifically on the spread and influence of 'dark propaganda' online, they consider the social elements such as online interactions, and the technological affordances that affect this. They also situate disinformation in the context of other media-related factors that might contribute to or drive the spread and influence of disinformation. However, their research focuses only on two specific case studies, (the United States[24] and Estonia). While they find differences between these cases, their research findings cannot necessarily be extrapolated to a wider geographical or situational sphere. Alongside these notions, it is useful to understand some of the reasons why people believe false content, and why they share it even when they know or suspect it is not true. A number of studies have been conducted concerning the psychology of belief, leading to the argument that behavioural sciences should play a key role in informing responses to disinformation (Lorenz-Spreen et al., 2020). Lewandowski looks specifically at conspiracy theories such as those around the coronavirus, claiming that in a crisis, people typically go through various stages of denial including not believing there is a crisis, blaming others for it, or not believing solutions will work, all typically leading to the support of these conspiracy theories (Cook et al., 2020).

In countries whose mainstream media is largely or fully controlled by government authorities, there is often a public distrust of such sources, particularly where this is linked with historical or current issues such as apartheid and corruption. In such countries, "radio trottoir" (literally, pavement radio) (Ellis, 1989) and other forms of underground media are often seen by the public as more trustworthy than official sources of information (Wasserman, 2020). Wasserman's study conducted in sub-Saharan Africa (Wasserman & Madrid-Morales, 2018) found low levels of trust in the media, a high degree of exposure to misinformation, and that people often contributed to its spread even with the knowledge

[24] The U.S. withdrew its membership from UNESCO in October 2017.

that facts were incorrect, to a much greater degree than United States (U.S.) citizens. This finding highlights the need to assess the extent to which strategies to counter disinformation should go beyond basic educational and media literacy strategies in order to tackle the root causes of mistrust.

The work of Wasserman and Ellis, among others, indicates that the reasons for knowingly sharing false information are likely to be connected with the notion of group allegiance. In other words, notions of truth are less important than notions of solidarity, and as long as a piece of information aligns with our world view, we often do not investigate its factuality. A study by Pennycook & Rand (2019) found distinct differences between people's ability to distinguish true from false information and their likelihood of sharing that information - in other words, it was not only the information they believed to be true that they said they would share. It is clear from all these findings that not only do "cognitive miserliness"[25] and cognitive bias play a part in our believing and sharing of false information, especially in an information-rich environment, but also that we are driven by heuristics such as social endorsement, and these elements should therefore be a factor in assessing responses to disinformation.

In order to respond effectively to disinformation, it is also important to understand some of the reasons why people are reluctant to change their opinions even when faced with evidence to the contrary. Hans Rosling discusses the notion that people typically have a number of negative misconceptions about the world (such as life expectancy in poorer countries, or the death rate from natural disasters), and even when faced with figures that disprove these, people struggle to accept them (Rosling, 2018). He blames this on three factors: fake nostalgia (a misremembering of the past as being better than it actually was); selective reporting by journalists (e.g. emphasising negative stories in accordance with traditional news values that prioritise exposure of suffering, corruption and wrongdoing in accordance with traditional news values); and a feeling that it is somehow inappropriate to talk about minor improvements during crises. The spread of disinformation often preys on and manipulates these beliefs, particularly where crises, conflicts and natural disasters are concerned. While Rosling encourages the notion of public education as a countermeasure, it remains a research gap to understand how effective this strategy is, especially given Rosling's own findings.

The "Ticks or It Didn't happen" report by Witness (Witness Media Lab, 2019) focuses on responses to disinformation from a primarily ethical viewpoint. Taking one of the core technologies for tracking image integrity ('point-of-capture' approaches at a camera level), the report reviews 14 dilemmas that are relevant since authenticity infrastructure is considered as a response to misinformation, disinformation and media manipulation. These dilemmas include technical dilemmas around access, as well as privacy, surveillance, government co-option, and concerns about setting overly-simplistic or hard-to-assess markers of credibility. The lens of the report is to use the framing of Collingridge's dilemmas (Collingridge, 1980) on the capacity to influence technological systems - and the challenge of doing that early enough to ensure they reflect human rights values, or risking being excluded once they are at scale. This lens is, however, also applicable to a range of technological approaches to disinformation, that may or may not prioritise freedom of expression or other human rights issues.

[25] Cognitive miserliness is the notion that we prefer to make easy decisions that align with our preconceptions, and may forget details (such as that the information had previously been debunked) https://firstdraftnews.org/latest/the-psychology-of-misinformation-why-were-vulnerable/

3.2.2 Vector focus

Alongside notions of persuasion and countering false beliefs, responses to disinformation also need to take into account the nature of the disinformation, and at whom it is aimed (as discussed above), but also the role of the conveyancing mechanism, or vector. These serve as intermediaries between the production and consumption of disinformation, enabling its circulation in various ways and at various scales. Knowledge about the patterns in this part of the cycle is critical for informing responses not only within transmission, but also in regard to strategies that target the initial production and subsequent consumption of disinformation.

There are three main mechanisms by which false content may be conveyed. First, disinformation may aim to disrupt or leverage the news media as a way to indirectly reach its targets, whether these be state or non-state actors. Captured media, compromised journalists, or weak capacities for verification constitute vulnerabilities that are exploited. Alternatively, disinformation may appear as a strategised (and often, but not necessarily, automated) exploitation and/or gaming of an internet platform to reach the public (i.e. targeting in part the nature of the business model and its reach). In other cases, disinformation is aimed primarily at the public for the purpose of onward dissemination, relying on its potential to trigger virality, using third parties to serve as peer-to-peer intermediaries to reach a bigger audience. In each case, responses need to target primarily the relevant mechanism (media, internet company, and public respectively).

3.2.3 Defending public values in a 'platform society'

While the news media and the public may serve as vectors for disinformation, this chapter now considers in more detail research into the role of internet communications companies (often referred to as 'platforms') as conduits, amplifiers and atomisers for disinformation. The rise of digital technologies has led to the increasing importance of data, with these companies emerging as new bastions of control and profit, having the facility to capture and manipulate enormous volumes of content- and, potentially, audiences. This in turn has led to the rise of dominant players (Srnicek, 2017), and it has important ramifications for the production, dissemination, and consumption of information and its reliability. An initiative by the NGO Public Knowledge, operating as https://misinfotrackingreport.com/, keeps pace with the policies and practices of a number of companies dealing with the challenges. Civil society movement Avaaz tracks the visible manifestations of disinformation narratives on specific themes, evaluating the performance of the companies in combatting such content.[26]

To some extent, the business models of digital platforms make them vulnerable as the conduits of disinformation, but there is also an argument that they are actually de facto enablers, or accomplices who turn a blind eye to the issue (Gillespie, 2017). Gillespie suggests a definition for the modern concept of (internet) platform as: "an architecture from which to speak or act, like a train platform or a political stage." However, like a growing number of researchers, he shuns the notion of 'platforms' because it tends to underplay the particular role of the companies involved. Gillespie points out that in reality, online platforms are not flat, open, passive spaces, but "intricate and multi-layered landscapes, with complex features above and dense warrens below." This suggests that such a complex structure influences how content is transmitted, and in ways that are not immediately open or straightforward. Instead, the nature of the online content that

[26] https://secure.avaaz.org/campaign/en/disinfo_hub/

users receive is shaped by algorithms, and can also change dramatically at the behest of those who have control of the design of the platform. The business model can enable bots and trolls to lurk beneath the surface and strike at unsuspecting victims or types of information, as well as enforce systemic biases such as decisions on what is allowed or not, and what might be a trending topic. That is one reason why the word 'platforms' is used sparingly in this report - instead, wherever feasible, the term 'internet communications companies' is used in preference.

Relevant to this issue are the financial gains to be made through the analysis of enormous amounts of data made available to companies which enable transmission or discovery of content. Zuboff (2019) has assessed how engagement is required from users, in order to produce this data, which is then monetised in the form of opportunities that are sold due to their ability to shape what she calls "behavioural futures". Reports from Ranking Digital Rights highlight that this business model leads to particular kinds of content becoming more widespread, including disinformation. By prioritising such content and recommending similar content, disinformation becomes increasingly linked with revenue for both platforms and the content providers, and the problem becomes circular (Maréchal & Biddle, 2020; Maréchal et al., 2020).

The book "The platform society: Public values in a connective world" (van Dijck et al., 2018) also offers an in-depth analysis of the role of these companies in shaping modern society. It focuses on public values in a world where social interaction is increasingly carried out on digital platforms, and investigates how these values might be safeguarded. Until recently, most companies have tended to evade acceptance of the social obligations related to their position as intermediaries of content, although this is beginning to change as pressure is put on them by authorities, especially European policymakers. While some companies have encouraged research into disinformation, there is reluctance to make their data available for this purpose. For example, Facebook has announced $2m for research into "Misinformation and Polarisation" with the proviso that "No data (Facebook, Messenger, Instagram, WhatsApp, etc.) will be provided to award recipients".

Another area ripe for further research in reference to the role of the internet communications companies' in combatting disinformation is the exploitation of 'data voids' (Golebiewski & Boyd 2019). Research being conducted at the time of writing, as part of a partnership between First Draft and researchers from the University of Sheffield, identified the particular problem posed by data voids during the COVID-19 pandemic. They found that when people searched for answers to questions about the causes, symptoms and treatments for coronavirus, the void created by the absence of verifiable answers to these questions (in part a product of the genuine scientific uncertainty associated with the onset a new virus; sometimes because of manipulated disclosure by authorities of statistical data) lent itself to exploitation by disinformation agents who filled the gap with spurious content: "If more speculation or misinformation exists around these terms than credible facts, then search engines often present that to people who, in the midst of a pandemic, may be in a desperate moment. This can lead to confusion, conspiracy theories, self-medication, stockpiling and overdoses." (Shane 2020) On the basis of preliminary findings, and recognising the role that social media sites now play as de facto search engines, the researchers called for a 'Google Trends' like tool to be developed for application to a range of social media sites including Facebook, Twitter, Instagram and Reddit, to enable easier and more transparent identification of disinformation being surfaced by such search activity.

The intersection between internet companies and news media companies as vectors for false content has also attracted some analysis. In particular, this highlights tensions between journalism and internet communication companies with respect to curatorial

efforts to counter disinformation and its viral distribution, the purveyors of which frequently target journalists and news publishers. These tensions have their roots in the 'frenemy' status of the relationship between these companies and news publishers (Ressa, 2019), which has been exacerbated by the collapse of traditional news business models, the erosion of historic gate-keeping roles, and the rise of 'platform power' (Bell & Owen, 2017).

The escalation of digital disinformation in the context of journalism's dependency on these social media networks for content distribution and engagement, and the platforms' encouragement of such dependency, have led to the phenomenon of 'platform capture'. Other examples of 'platform capture' include the ways in which efforts to curtail disinformation can backfire, such as WhatsApp's change in terms of service in 2019 which negatively affected the media's ability to use the technology to counter disinformation (Posetti et al., 2019b).

Traditional journalism commits to a set of news values (Galtung and Ruge, 1965) that include accuracy, verification, and public interest, but this is potentially orthogonal to the values of digital platforms which typically include innovation and peer-to-peer connectivity (Wilding et al., 2018), not to mention monetisation at the expense of editorial standards. As Foer (2017) indicates, dependence of the news media on the values of the digital platforms, means that their intensified quest to go viral risks superseding the quest for truth. This problem is further exacerbated by algorithms for the optimisation, dissemination and even production of news (Wilding et al., 2018) as well as search engine optimisation.[27] In addition, audience engagement has become a core driver, resulting in a change in news production towards a "softer" form of news (Hanusch, 2017) that is shorter, more visual, and more emotive (Kalogeropoulos et al., 2016). Added to this, 'content farms' are producing or recycling questionable low-quality content with dubious factuality but which are optimised for engagement.

The digital transformation of journalism is ongoing - change is now regarded as a perpetual - therefore, it is important that research keeps pace with the associated challenges and opportunities relevant to the production, dissemination and amplification of disinformation in the 21st century news ecosystem (Ireton & Posetti 2018).

An assessment of the internet and news media vectors, and the relationship between them, are discussed in detail in Chapter 6.1 of this report.

3.2.4 Policy-driven approaches to studying disinformation

The COVID-19 crisis prompted a range of studies with a view to developing policy responses, including by UNESCO (Posetti & Bontcheva, 2020a and 2020b) and the OECD (2020). The OECD study used the Wardle and Derakhshan (2017) framework to identify four governance responses to disinformation: identifying and debunking; civic and media initiatives; communications strategies; and regulatory measures. Particular attention was focused on public communication with the message that "Strategic and transparent communication should be among the first lines of action for public institutions at all levels".

The LSE's *Tackling the Information Crisis report* (LSE, 2018) explains how changes in the UK media system have resulted in what it calls an information crisis. It depicts this as being

[27] https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/08/Follow-the-Money-3-Aug.pdf

manifested in 'five giant evils' among the UK public – confusion, cynicism, fragmentation, irresponsibility and apathy. It also summarises a number of UK policy responses, including UK parliamentary inquiries; UK government initiatives including among other things the Digital Charter (UK DCMS & Rt Hon Matt Hancock, 2018b), a white paper on new laws to make social media safer (UK DCMS, Home Office, Rt Hon Matt Hancock & Rt Hon Sajid Javid, 2018a), and the new DSTL Artificial Intelligence Lab in Porton Down, whose remit includes "countering fake news" (UK MOD et al., 2018); institutional responses such as those by Ofcom (2018b) and the Commission on Fake News, and the teaching of critical literacy skills in schools (National Literacy Trust, 2018). While the report provides a detailed coverage of policy responses to disinformation, it focuses primarily on recommendations and recent initiatives, but research is still needed on analysing the outcome and impact of these.

Launched in November 2018, the Information Warfare Working Group[28] at Stanford University, comprised of an interdisciplinary group of researchers at the Center for International Security and Cooperation at the Freeman Spogli Institute and the Hoover Institution, aims to "advance our understanding of the psychological, organizational, legal, technical, and information security aspects of information warfare", working towards producing a set of policy recommendations for countering foreign disinformation threats. They have so far produced a number of white papers and reports. The work comprises research from many different disciplines and foci, while at the same time it focuses rather narrowly on political aspects of disinformation in the U.S..

Other important resources at the European policy level include a study commissioned by the European Parliamentary Research Service investigating the effects of disinformation initiatives on freedom of expression and media pluralism (Marsden & Meyer, 2019), as well as the work of the High Level Expert Group (HLEG) on 'Fake News' and Online Disinformation (Buning et al., 2018).

The first of these reports examines the tradeoffs between the application of automated (AI) techniques to counter disinformation, focusing mainly on ways in which EU legislation can be used to drive the design of these technologies in a way that does not restrict freedom of expression unnecessarily, and which maximises transparency and accountability. It thus focuses primarily on technological and legislative responses to disinformation, and raises concerns over the nature of current legislation that might restrict freedom of expression, concluding that there is a lack of policy research in this area, and that single solutions, particularly those which focus primarily on technological responses, are insufficient. In a similar vein, the HLEG report provides a policy-driven perspective on disinformation, advising against simplistic solutions and encouraging holistic solutions promoting maximum transparency, information literacy and empowerment, and suggesting a mixture of short- and long-term actions.

Both these reports thus focus specifically on European policy issues, and thus do not consider how this might be translated beyond these boundaries. Indeed, a major research gap in all the existing policy-driven reports is that each proposes their own set of strategies but it is unclear how to proceed from this to an overarching set of responses, even though disinformation clearly does not respect geo-political boundaries.

A group of experts from the University of Pennsylvania have produced a report titled "*Freedom and accountability. A transatlantic framework for moderating speech online*" (Annenberg Public Policy Center, 2020). This document states that: "Through a freedom-

[28] https://cisac.fsi.stanford.edu/content/information-warfare-working-group

of-expression lens, we analyzed a host of company practices and specific laws and policy proposals, gathering best practices from these deep dives to provide thoughtful contributions to regulatory framework discussions underway in Europe and North America." To deal with online problems, including disinformation, the report proposes that States should regulate internet companies on the basis of compulsory transparency provisions, and that there is also regulatory oversight to "hold platforms to their promises". For the internet companies themselves, the report suggests a three-tier disclosure structure, effective redress mechanisms, and prioritisation of addressing online behaviour by "bad actors" before addressing content itself.

3.2.5 Practice-relevant studies and resources

The Digital News Report[29] of 40 markets from the Reuters Institute for the Study of Journalism documents the role that internet companies are now playing in the distribution of both information and what the report's authors call "misinformation". It points out that audiences can "also arrive at misinformation (as they arrive at much else) side-ways via search engines, social media, or other forms of distributed discovery". The 2018 report in the series examined variations in exposure and concern, and different beliefs about remedies to false content online. The Institute has also researched types, sources, and claims of COVID-19 misinformation (Brennan et al., 2020), and mapped disinformation responses from three Global South news organisations re-conceptualising themselves as 'frontline defenders' in the 'disinformation war' (Posetti et al., 2019a; Posetti et al., 2019b). These latter reports identify enhanced methods of investigative reporting (including big data and network analysis), advanced audience engagement techniques (such as collaborative responses to surfacing and debunking disinformation), and 'advocacy' or 'activist' models of journalism (that involve actively campaigning against disinformation vectors, or providing digital media literacy training to their communities) as methods of responding to the disinformation crisis.

Jigsaw (an arm of Google) has produced what they term a **visualisation of disinformation campaigns** around the world, supporting their theory that "understanding how disinformation campaigns operate is one of the first steps to countering them online".[30] They state that this visualisation is based on the Atlantic Council's DFRLab research and reflects their perspectives in the characterisation. Additionally, they note that their visualisation is primarily based on open source, English-language press reporting of campaigns which appear to target the West. These kinds of visualisation provide an interesting overview, despite geographic limitation, but risk conflating very different kinds of disinformation.

A noteworthy set of practical resources pertaining to disinformation includes some of those discussed in chapters 7.1 (focused on normative and ethical responses) and 7.3 (educational responses), which not only support practical skills, but also investigate underlying theories and trends. The UNESCO handbook *Journalism, 'Fake News' and Disinformation* (Ireton & Posetti, 2018), is a research-based educational resource aimed at journalists and news organisations confronting disinformation, with an emphasis on freedom of expression issues. In addition to its role as a set of resources to support journalism education, it also explores the nature of journalism with respect to trust, as well as the structural challenges that have enabled viral disinformation to flourish, and the conduits of information disorder such as digital technology and social media, and it

29 http://www.digitalnewsreport.org/
30 https://jigsaw.google.com/the-current/disinformation/dataviz/

describes the targeting of journalists and their sources in the context of disinformation campaigns. The book offers a framework for understanding independent, critical journalism as a mechanism for combatting disinformation. It also provides models for responding innovatively to the challenges of disinformation as they impact on journalism and audiences. Among other resources of this kind is the *Verification Handbook for Disinformation and Media Manipulation* produced by the European Journalism Centre (Silverman, 2020).

Examples of resources focusing on the public include the UNESCO MIL Digital Toolkit[31] comprising MOOCs on Media and Information Literacy in several languages, and the International Center for Journalists' (ICFJ) learning module on the history of disinformation (Posetti & Matthews, 2018). One important gap in a number of these toolkits and programmes is a focus on the wider representation of 'data', including privacy and profiling issues, and more generally how data is collected and used by online platform providers, as discussed earlier in this chapter. The experience of the NGO 5Rights, for example, has shown that when children understand these concepts, their overall information literacy also improves. However, many skills-based approaches to countering disinformation only focus on the basic concepts of verification of the immediate sources without considering these wider foundational aspects.[32]

Finally, a handbook for government communicators on countering information influence activities has been produced by the Swedish Civil Contingencies Agency (MSB, 2020).

31 https://en.unesco.org/MILCLICKS
32 https://5rightsfoundation.com/uploads/digital-childhood---final-report.pdf

3.3 Current research gaps

As has been indicated, there is a plethora of research on disinformation and approaches to countering it, both from a theoretical and practical standpoint. However, there is an apparent disconnect between academic research, journalistic investigations, and studies commissioned by civil society and intergovernmental organisations. Additionally, actual collaboration between these sectors appears to be infrequent. The initiatives and publications mentioned have been produced in an ad-hoc manner, and are disparately located, making it difficult to track, analyse, and synthesise them in a coherent way. For example, cross-institutional study of the relationship between the technological/business logic and the realm of company and state policies is still weak, as will be discussed further in chapter 6.2.

The impact of most of the responses counteracting disinformation has also not been studied sufficiently. While some research has investigated which groups (such as elderly people) are particularly susceptible to both believing and sharing disinformation (Carey et al., 2020; Guess et al., 2019), there have been few responses directly aimed at vulnerable groups, and there is a dearth of empirical assessment of these, with exceptions like Humprecht et al. (2020), although with limited geographical focus. Linked to this, methods of countering disinformation have also not sufficiently covered notions of group allegiance and distrust in authority, which require a different outlook and more fundamental issues to be addressed.

Finally, while there is a growing body of research, software, training and resource development focused on tackling disinformation, there is a comparative absence of that which focuses on disinformation in the light of human rights, freedom of expression, and the burgeoning access to - and use of - broadband technology worldwide.

Below is a further analysis covering some particular areas where important gaps have been identified.

Addressing distinctions and connections between realms of disinformation

In terms of frameworks, much published research does not make a clear distinction between novel kinds of disinformation (for example, deepfakes) and those with much older histories (such as notions of information influence, which overlap with disinformation as discussed above). Others apply only in specific contexts, such as political disinformation, or may have limited applicability to non-Western nations (Brooking et al., 2020). A number of frameworks also view disinformation not only in a political light, but also focus primarily upon foreign influence, and thus do not address the numerous issues related to domestic disinformation, such as that pertaining to health crises, issues of migration, and disaster communications.

On the other hand, there are separate specific studies around such issues, as witnessed by the latest efforts to map disinformation around the COVID-19 pandemic and to implement counter-strategies, discussed in more detail in the following chapters. In general, the effect of the pandemic has been to ramp up public awareness of disinformation, and educational efforts promoted by both state and non-state actors (governments, internet communications companies, media companies, etc.). COVID-19

has provided a clear case where the effects and harms of disinformation can be easily seen, thereby elevating its importance and dangers in the public's eyes, and may lead to increased research, such as the initiative of the World Health Organisation (WHO) to explore an interdisciplinary field of "infodemiology" study, which has relevance to fields outside of health.[33]

Data availability for research

In terms of understanding the nature of disinformation, its dissemination and counter-activities, the issue of the lack of transparency of algorithms behind social media platforms and issues with access to their data is a serious hindrance, as discussed in chapter 4.2. Quantification of disinformation online relies on selective disclosure by the companies and what is contained in their transparency reports, without researchers having access to original data.

There is evidence, from external studies, about instances of disinformation pieced together through content analysis techniques. One snapshot study said it found that one in four popular YouTube coronavirus videos contained misinformation.[34] This research analysed 69 of the most widely-viewed English language videos from a single day in March 2020 and found 19 contained non-factual information, garnering more than 62 million views. In another study, an analysis of more than 1300 Facebook pages with nearly 100 million followers produced a network map showing that while anti-vaccine pages have fewer followers than pro-vaccine pages, they are more clustered and faster growing, and increasingly more connected to other pages.[35]

Such findings signal the importance of assessing patterns of disinformation online, and they also show what can be done even without data disclosed by the internet companies.

Nevertheless, most research into disinformation is limited by being conducted without access to the complete data sets from the internet communications companies. This leads to a lack of depth in their analysis, and studies are also typically carried out only on a selected platform (frequently Twitter with its volume of open and public data), rather than cross-platform. Messaging apps are rarely considered due to their closed nature. Social media companies present a number of obstacles to independent research by cutting access to APIs by which researchers can collect relevant data, mirroring to some extent the problems with search engine research, where only those with direct relationships with the major search companies can work effectively (Walker et al., 2019). For instance, it is hard to know specifics when users or messages are removed by the provider (or when the user retracts the information themselves). While these platforms do offer a selected group of academic researchers to access such data via research grants[36] by means of tools such as Crowdtangle[37], at the time of writing this was limited in scope and included restrictions on the kinds of research that could be done. In the light of COVID-19, Crowdtangle launched (in March 2020) more than 100 publicly available LiveDisplays enabling researchers to investigate issues such as the spread of information about the pandemic on social media, nevertheless this still provides only a restricted set of data.

33 https://www.who.int/docs/default-source/epi-win/infodemic-management/infodemiology-scientific-conference-booklet.pdf?sfvrsn=179de76a_4
34 https://www.nbcnews.com/health/health-news/live-blog/2020-05-13-coronavirus-news-n1205916/ncrd1206486#liveBlogHeader
35 https://www.nature.com/articles/s41586-020-2281-1
36 https://about.fb.com/news/2019/04/election-research-grants/
37 https://www.crowdtangle.com/

The report of the Annenberg Public Policy Centre (2020), cited above, argues that transparency enables governments to develop evidence-based policies for oversight of internet companies, and pushes firms to examine problems they would not otherwise address, and thus empowers citizens. This insight points to the value of companies providing much greater access to data. Companies are understandably sensitive about providing data for reasons of commercial secrecy as well as avoiding data compromises, as occurred during the Cambridge Analytica experience. Against this background, MacCarthy (2020) has proposed the nuance of a tiered model for access to company data, distinguishing different levels that could be availed to the public, vetted researchers, and regulators.

The consumption and response to disinformation

Studies in **user behaviour and perception** are still lacking, not least in regard to the relationship between the impacts of disinformation and of news. For example, even when faced with a diverse selection, people tend to choose news articles that are most aligned with their own beliefs (Kelly Garrett, 2009) - through user-driven customisation or selective exposure, reinforced by predictive algorithms. Nevertheless, little work has been carried out on assessing its actual effect. This has important ramifications for disinformation with respect to issues of propaganda or dangerous health-related beliefs such as those promoted by anti-vaccination supporters. The implications of such selective exposure are of increasing concern, since they can enhance social fragmentation, mirroring or amplifying enduring cleavages, thereby also reinforcing pre-existing opinions and perceptual biases. The correlation between exposure to misinformation and effects on offline behaviour also requires further investigation, such as the relationship between misinformation, fear, panic, and unselfish and irrational behaviour (see e.g. Osmundsen et al., 2020).

Competing notions currently exist around the extent and effect of exposure to different viewpoints on one's ideological perspectives. On the one hand, the increasing use of social media and personalised news acts as a 'filter bubble' or 'echo chamber', reinforcing existing beliefs and increasing ideological segregation. However, there is a growing body of empirical research arguing that the effect of filter bubbles has been overstated (e.g. Dubois & Blank, 2018; Guess et al., 2018a), and that only a small subset of people tend to have heavily skewed media consumption (Gentzkow & Shapiro, 2011; Flaxman et al., 2016), something which extends also to misinformation (Guess et al., 2018b). Others posit that the increasing availability of information, coupled with the consequent greater diversity of the information consumed, actually widens the range of news sources to which people are exposed (Fletcher & Nielsen, 2018). Another study showed that even users of very different political backgrounds were typically exposed to very similar sets of political news (Nechushtai & Lewis, 2018), contradicting theories about the filter bubble effects of news personalisation.

What is unclear is what effect widening the exposure to different viewpoints might have on issues of ideological partisanship. Understanding and measuring ideological diversity from big social data, and the influences on ideological perspectives that might be brought about by exposure to such diversity, would all lead to improved understanding of the effect of disinformation and counter-content such as fact-checking and verified journalistic news. Large-scale user studies would be needed in order to better understand how people evaluate the truth and reliability of information – both from a practical perspective and from a psychological perspective. Similarly, studies targeting users' behaviour in relation to engagement with, and redistribution of, credible, verified

information - such as that produced by independent news publishers and journalists - could provide insight.

Several attempts have been made to mitigate the effect of bias in information systems to support an unfiltered process of opinion formation. Some have focused on making users aware of bias by providing alerts (Epstein & Robertson, 2015), visualising reading behaviour and bias (Munson et al., 2013), or pointing to web pages with different opinions from the current one. Others rely on visualisations to support diversity in web activities (Graells-Garrido et al., 2016), recommendation systems (Tsai and Brusilovsky, 2018), and search results (Verberne, 2018). Some focus on algorithm transparency by explaining how filtering works and enabling the user to control the algorithm and thus their filter bubbles (Nagulendra & Vassileva, 2014). Others try to break potential filter bubbles through software design and user interfaces (Bozdag, E., & van den Hoven, J., 2015). However, success in all of these approaches is rather limited (Faridani, 2010; Liao & Fu, 2013), and more studies are clearly needed to better understand online news consumption patterns and habits, such as how people navigate the constantly changing environments to select which news they decide to read (Swart et al., 2017).

The changing technological and institutional infoscape

It can be noted that many of the responses to disinformation described in this report are still quite new and not yet widely implemented. This may be because the technologies are still being developed or improved, because they are only adopted by a small minority, or for other reasons such as legal and ethical issues which need to be resolved. For example, when credibility and labelling approaches are not widely used, this not only has clear limitations on their effectiveness, but also on the understanding of their potential. It is simply not known if they will be successful until they are rolled out more widely. There are also potentially serious implications if they are applied at scale, as detailed in the 'Ticks or It Didn't Happen' report by Witness (Witness Media Lab, 2019). This illustrates Collingridge's dilemma (Collingridge, 1980), which essentially posits that the social consequences of technology often cannot be predicted until the technology has already been developed, at which point it is often too late, or at least much more difficult to change. Neither Collingridge nor the Witness report suggest that these challenges cannot be overcome, but focus on early consideration of scenarios, as well as flexibility of approach in order to deal with them.

Related to this, evaluation of many of the technologies proposed to counter disinformation is still lacking, and furthermore little discussed. It is not always even clear how effective some of the methodologies are in principle, such as the notion of fact checking, since research has shown that the reach of fact-checked material is often very different from the reach of the disinformation itself, and indeed, instances of a "backfire effect" have been witnessed where corrections can sometimes even increase misperceptions (Nyhan, 2012). More research could help in evaluating the effect not only of the technologies but also their underlying theories of change, which may be based on false or misguided assumptions. Further discussion of this is presented in Section 4.1.

International representativeness in research

The Global South in particular has typically been under-represented in terms of research focus. Examples include Chaturvedi's study of India (Chaturvedi, 2016); Kaur et al.'s study of Asia and the Pacific (Kaur et al., 2018); recent reports of joint research between FullFact, Chequeado and Africacheck focusing on Argentina, South Africa and Nigeria[38], and the Oxtech report on anti-disinformation initiatives, which uses examples from 19 countries on four continents.[39] Reports from a policymaker's perspective, in particular, are almost exclusively focused on Europe and North America. That is a clear gap that this study aims to address, partly in the hope that it will trigger investment in future action-oriented research.

All this highlights the value of a large-scale global study such as this one, which collates the multiplicity of disinformation responses from a variety of perspectives, and incorporates the needs and challenges of culturally distinct geographical regions.

Human-rights dimension

Few conceptual frameworks or other literature really focus on the critical problem of ensuring a balance between protecting freedom of expression and upholding notions of truth, against disinformation, although this is connected implicitly with some of the discussions in this chapter around the internet communications companies, as well as around journalistic integrity. Meanwhile, regulating speech on social media in an attempt to prevent disinformation clearly has ethical and policy implications that intersect with freedom of expression, as does the passage of legislation creating 'fake news' laws that represents a significant threat to press freedom. The EU Code of Practice on Disinformation (European Commission, 2018c) has recently been criticised for theoretically allowing, and even incentivising, restrictions on the freedom of speech that are claimed to be technically lawful (Kuczerawy, 2019). Kuczerawy voices concerns that enlisting private platforms to suppress certain online content that is not illegal may have unintended consequences, and argues that it is difficult to "factually assess the impact of the Code on the exercise of the right to freedom of expression". In countries outside the EU, where less stringent regulations may apply, there is the potential for greater concerns of this nature. These issues are discussed more fully later in this report, in particular in the discussions of legislative responses to disinformation in Chapter 5.1, as well as in the discussions of policy responses in Chapter 5.2, since both these kinds of responses must deal with this exact issue.

38 https://fullfact.org/research/

39 https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/A-Report-of-Anti-Disinformation-Initiatives

3.4 Novel contributions of this study

Having situated this study within the context of existing theoretical frameworks and previous research, and having identified the gaps in current research on the topic of disinformation, this section highlights the specific novel contributions presented here.

Firstly, this study has sought to adopt a global focus, while many of the reports cited above have largely focused on particular countries or continents and a great amount of research has centred on the UK, U.S. and/or European situations. This partly reflects the fact that these geographical regions are highly active in responses to disinformation, and that they represent the location of the majority of researchers and funding for investigating the topic. Further, the dominant disinformation sources under examination in other reports have been limited to English language content.

By contrast, this report has sought to focus on issues and initiatives worldwide, including those from Africa, Australia, Central and Eastern Europe, Latin America and Asia. For example, this has helped reveal that some journalistic responses to disinformation rely on having certain technological requirements, or are difficult to adopt for those in conflict situations (such as when reporters need to maintain anonymity and cannot use certain point-of-capture tools for photos and videos as a result). Below, we discuss how and why particular responses may be difficult for actors in certain countries and situations, which are not necessarily considered by those in Western Europe and the U.S..

The authors of this report are of diverse ethnic and regional backgrounds, they speak a variety of languages and they possess specific knowledge about situations in different parts of the world. They also come from a range of disciplinary backgrounds. The research team includes members from both academia and industry, with a mixture of computer scientists, journalists, social scientists (including those with a journalism studies and political science background), and specialists in international human rights with an emphasis on freedom of expression. This leads to an approach which addresses a range of perspectives and is closely tied to both practice and impact. There is thus also a focus on technical responses such as the use of AI, in addition to educational responses, responses from the journalism sector, and responses from the industrial technology sector.

This report is also novel because it puts the main focus specifically on **responses** to disinformation. As discussed above, other notable reports focus on dilemmas (Witness Media Lab, 2019), policy implications (e.g. LSE, 2018; Annenberg Public Policy Center, 2020), political implications (e.g., Marsden & Meyer, 2019; Pamment et al., 2018), and significance for (as well as responses from) journalism (Ireton & Posetti, 2018; Posetti et al., 2019a). Furthermore, this report addresses the entire spectrum of disinformation responses, rather than focusing on a specific type such as political disinformation campaigns (Brooking et al., 2020) or issues with access to company data and how this affects academic research (Walker et al., 2019).

A further novel angle of this study is that the problem of disinformation is systematically addressed in the light of freedom of expression challenges, with implications for press freedom such as in legislative responses, among others.

The typology of responses that this study has developed also breaks down the problem of disinformation in a new way. It examines each response from a variety of perspectives, looking beyond the what and how to issues such as "Who is funding these responses (and the implications thereof)?", "What are the strengths and weaknesses of them?", and "What is the theory of change on which they are based?" This approach provides additional insight into the assumptions upon which the responses rest, and the extent to which they integrate monitoring and evaluation into their activities.

Identification Responses

4

4.1 Monitoring and fact-checking responses

Authors: Denis Teyssou, Julie Posetti and Sam Gregory

This chapter maps out and systematises fact-checking responses as part of monitoring disinformation (alongside investigative responses to disinformation, as covered in chapter 4.2). Here, the emphasis is on fact-checking responses which are global, regional and national in scope in a wide range of the countries and languages, and they can be either independent operations or affiliated with news organisations. The way these efforts engage in countering disinformation is also described in this chapter.

Definitions

The discipline of verification has been described as the essence of journalism (Kovach & Rosenstiel, 2001). Verification is an editorial technique used by journalists and by independent fact-checkers to verify the accuracy of a statement, and/or documents and other artifacts, as well as the platforms and identities (human and digital) of those producing or transmitting content. But there are distinctions to be drawn between verification and fact-checking (Silverman et al., 2014):

- **Verification** *is a discipline that lies at the heart of journalism, and that is increasingly being practiced and applied by other professions.*
- **Fact checking** *is a specific application of verification - both within journalism* [*and by other organisations, including NGOs*]. *In this respect, verification is a fundamental practice that enables fact checking.*

Increasingly, fact-checking also involves a process of proactive de-bunking - i.e publishing debunks to demonstrate falsehoods, and often by setting out the systematic process involved in reaching this conclusion.

4.1.1 What and who do they target?

Fact-checking responses consist of applying verification not only to the process of journalistic work (and its outputs), but also to third-party claims, statements and datasets circulating outside the legacy media sphere, especially on social networks.

Verifying the authenticity of an actor, institution or a social media account is where fact-checking begins to feed into investigative responses to disinformation (see Chapter 4.2). Identification responses, like monitoring and fact-checking, underpin the investigations into the origins and spread of disinformation, contributing to the evidence-base upon which other types of disinformation responses depend.

Specific examples will be provided in section 4.1.4 below.

4.1.2 Who do monitoring and fact-checking responses try to help?

The usefulness of fact-checking for internet communications companies[40] enables them to identify disinformation and develop responses that reduce or remove its visibility and/or credibility. Checking also helps governments and international organisations to decide what, when and whether action needs to be taken - for instance, launching policy or practical initiatives like targeted counter-disinformation campaigns. Finally, published fact-checks provide a useful source of authoritative information for citizens.

4.1.3 What output do they publish?

This response publishes its findings - what was checked, how, and what the status is in terms of validity or falsity, indeterminate or other (e.g. opinion - which is not fact-checkable per se, although where it is justified on the basis of purported facts, these aspects are prima facie checkable concerning the extent to which such 'facts' are false or misleading). It is recognised that published fact checks tend to attract fewer user shares on social media than the viral disinformation they are debunking (Shin & Thorsen, 2017). There is also some concern that drawing attention to falsehoods can help amplify them. Nevertheless, the operating assumption is that the work of verification and debunking remains essential as a means for surfacing truth and for holding individuals, public figures, institutions and the media accountable for inaccurate claims (Sippitt, 2020; Friedman, 2020; Qui, 2020).

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

a. Global responses

First Draft

One of the early global initiatives focused on social media content verification at the international level is the non-profit coalition First Draft, registered in the UK since June 2015. The aim of First Draft at its establishment was to provide practical and ethical guidance to the news media on identifying, verifying and publishing content that is sourced from the social web, especially in breaking news contexts.

In September 2016, the original nucleus of nine partners (BellingCat, Dig Deeper, Emergent.info, EyeWitness Media Hub, Google News Initiative, Meedan, Reported.ly, Storyful, and VerificationJunkie) expanded to an international Partner Network of media organisations, academics, social network platforms and civil society organisations. At the same time, First Draft joined ProPublica's project ElectionLand, which aimed to identify and track voters' encounters with misinformation and disinformation during the 2016 U.S. presidential election. They worked collaboratively on this project with students from 13 journalism schools who were trained in social newsgathering and verification techniques. Electionland was financially supported by Google News Lab and U.S. philanthropist Craig Newmark.

[40] https://www.disinfo.eu/resources/covid-19/platforms-responses-to-covid-19-mis-and-disinformation;

Next, First Draft launched several collaborative election-monitoring programs in France, the United Kingdom, Germany, Brazil, and Nigeria. The resulting news media and fact-checking coalition, known as CrossCheck, monitors rumours being spread within these countries, and publishes debunks of false information in order to give voters the means to reach conclusions by themselves without being misdirected by disinformation. (For more on election-targeted responses, see section 5.3).

In 2020, First Draft was expanding operations in Argentina, Australia, Canada, Indonesia, South Africa, Spain and Uruguay, and aiming to coordinate a cross-border project to investigate misinformation tactics and trends in Europe beyond election periods.

Apart from founding partner Google News Initiative, First Draft has also obtained grants and donations from many philanthropic foundations as well as support from the Facebook Journalism Project and Twitter. After briefly joining the Shorenstein Center for Media, Politics and Public Policy at Harvard's Kennedy School in October 2017, First Draft is now operating independently again, primarily relying on funding from internet communications companies. More First Draft collaborative initiatives around elections are detailed in section 5.3.

International Fact Checking Network (IFCN)

The International Fact-Checking Network (IFCN, 2019a) was launched in September 2015 as a business unit within the non-profit journalism school Poynter Institute for Media Studies, based in St. Petersburg, Florida, U.S.. The Institute, which owns the *Tampa Bay Times*, launched IFCN to bring together fact-checkers worldwide and to promote good practices and knowledge exchange in the field.

The IFCN's mission is to monitor trends, formats and policy-making about fact-checking worldwide, to publish regular articles about fact-checking, to promote training - both in person and online - as well as ensuring basic standards through the fact-checkers' code of principles. On August 6th, 2020, IFCN had 79 verified active signatories of its code of principles, 14 verified signatories under renewal (IFCN, 2020d). The map below shows the geographic distribution of the signatories. Some of them are fact-checking both in their homelands and across international borders.

A verification process is important because it is possible that in this contested terrain that flawed, or even fake, fact-checking initiatives can exploit the label for purposes far removed from challenging falsehoods.

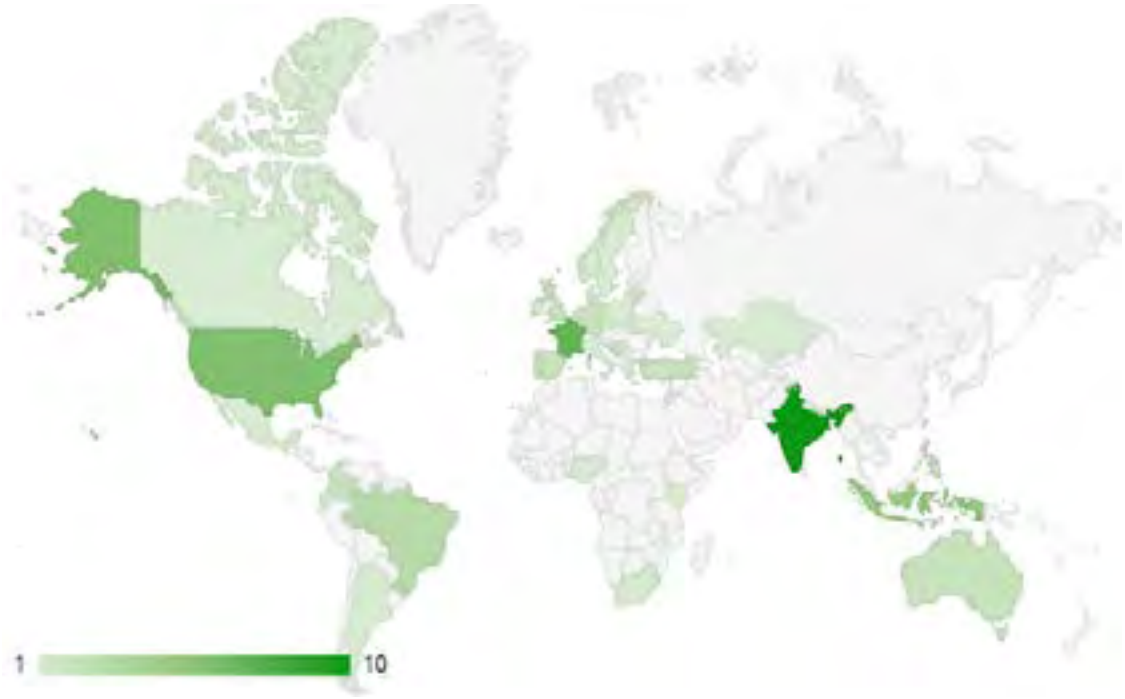


Figure 4. *A geographical map of the IFCN signatories* **(***67 verified active and 14 under renewal in early 2020***)**

Signatories must abide by five commitments (IFCN, 2019c):

1. A commitment to Nonpartisanship and Fairness
2. Transparency of Sources
3. Transparency of Funding and Organisation
4. Transparency of Methodology
5. Open and Honest Corrections Policy

This code of principles was launched in September 2016, one year after the birth of the IFCN. In 2017, the IFCN introduced an application and vetting process following the announcement by Facebook that being a signatory to this code is a minimum condition for being accepted as a third-party fact-checker for the company.[41]

Transparency, often presented in media studies literature as a new ethical tenet of journalism, plays an important role in these commitments. This intersects with the emergence of transparency among bloggers and early fact-checkers as a necessary or natural alternative to the professional journalistic ideal of objectivity (Graves, 2013). It builds on an idea espoused by philosopher David Weinberger in 2009: "transparency is the new objectivity" (Weinberger, 2009). The notion of transparency and its connection to trust in credible journalism is now widely embedded as a norm within both fact-checking operations and professional journalism. The transparency afforded by published explanations of verification and fact-checking processes can make the work more defensible against claims of bias or inaccuracy because the evidentiary base of the fact-checking exercise is laid bare.

[41] See chapters 4.2 and 7.1 for further discussion

The IFCN organises a yearly international conference (Global Fact) promoting collaborative efforts between fact-checkers all over the World. The 2019 edition, Global Fact 6, was staged in Cape Town, South Africa, with more than 250 participants representing 55 countries and 146 active organisations. Global Fact 7, which was due to be held in Oslo, Norway, in June 2020, was ultimately held online due to the COVID-19 pandemic. The Network also funds annual fellowships, a Fact Forward Fund, a Fact-Checking Innovation Initiative and a crowdfunding match program. Finally IFCN advocates for a global expansion of fact-checking efforts, including through an annual International Fact-Checking Day, every April 2.

IFCN has received funding from the Arthur M. Blank Family Foundation, the Duke Reporters' Lab, the Bill & Melinda Gates Foundation, Google, the National Endowment for Democracy, the Omidyar Network, the Open Society Foundations and the Park Foundation (IFCN, 2019d).

Duke University Reporter's Lab database

The Reporters' Lab is a centre for journalism research in the Sanford School of Public Policy at Duke University in the U.S.. One of its main projects has been to create a worldwide database[42] of the main fact-checking operations, active or inactive, and therefore to document the rise of the fact-checking sector, country by country.

Aside from a geographical mashup displaying all fact-checking organisations, the database allows the user to browse the content by continents and countries and it is regularly updated. Criteria to add new fact-checking sites include non-partisanship, an emphasis on reviewing fulfilment of political promises (e.g. party manifestos during elections), transparency about sources and methods, transparency about funding and affiliations, and a primary mission being news and information. As of April 2020, the Reporters' Lab database included 237 active sites and 91 inactive worldwide in 78 countries.

Figure 5. *A view of the Duke University Reporters' Lab fact-checking database*

42 https://reporterslab.org/fact-checking/

Facebook Third-Party Fact Checking network

Internet communications companies typically employ internal or external fact-checking processes, which inform their curatorial responses, e.g. removal, demotion, or hiding of posts. These are described in more detail in Chapter 6.1.

Among the companies' systems, Facebook merits attention as the only large-scale international "third party verification" programme among the internet communications companies, which was launched shortly after the 2016 U.S. presidential election (Zuckerberg, 2016a). Announcing the project on his own Facebook page, CEO Mark Zuckerberg stated that Facebook was taking "misinformation seriously," and acknowledged that there were many respected fact-checking organisations that his company was planning to learn from. Previously, he had stated that more than 99% of what people see on Facebook is authentic (Zuckerberg, 2016b). His announcement of the third party fact-checking initiative was widely interpreted as an attempt to counter criticism of the company's lack of intervention to stem the spread of disinformation during the 2016 U.S. presidential election campaign.

One month after the U.S. election in 2016, Facebook announced the launch of a programme to work with third-party fact checking organisations who were signatories of the Poynter Institute's International Fact Checking Network (IFCN) Code of Principles (Mosseri, 2016). The Facebook third party fact-checking program sub-contracts organisations to review and rate the accuracy of content, including stories and non-political advertising (See discussion below about the limitations applied to fact-checking political content under this program).

Once a story is rated as false by these fact-checking partners, Facebook shows it lower in the 'Newsfeed' unless it is revised by Facebook in light of their policies, processes and/or payments associated with the contracts under which the fact-checking organisations operate (Pasternack, 2020). (Generally opinion content, and generally certain categories of political advertising and political speech from politicians, political parties and affiliates are excluded). On Instagram, Facebook makes content flagged under this program harder to find by filtering it from Explore and hashtag pages, and downranking it in the feed. In addition, content across Facebook and Instagram that has been rated false or partly false is prominently labelled[43] so people can better decide for themselves what to read, trust, and share. These labels are shown on top of false and partly false photos and videos, including on top of 'Stories' content on Instagram, and link out to the assessment from the fact-checker.[44]

Prior to December 2017, if a fact-checking organisation identified a story as false (or 'fake' according to Facebook's protocol), they reported it to Facebook and it was flagged as disputed, with a link to the corresponding article (on fact-checker's website) explaining why.

According to Facebook, this limits the visibility of such posts by 80% (Lyons, 2018a) and therefore helps contain its spread. However, this can take up to three days after the content is first published (Silverman, 2017b). Facebook says that it also uses the information from fact-checkers in order to improve its technology to identify false content faster. Further assessment of labelling can be found in chapter 7.3.

[43] https://about.fb.com/wp-content/uploads/2020/06/Elections-Fact-Sheet.pdf
[44] https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

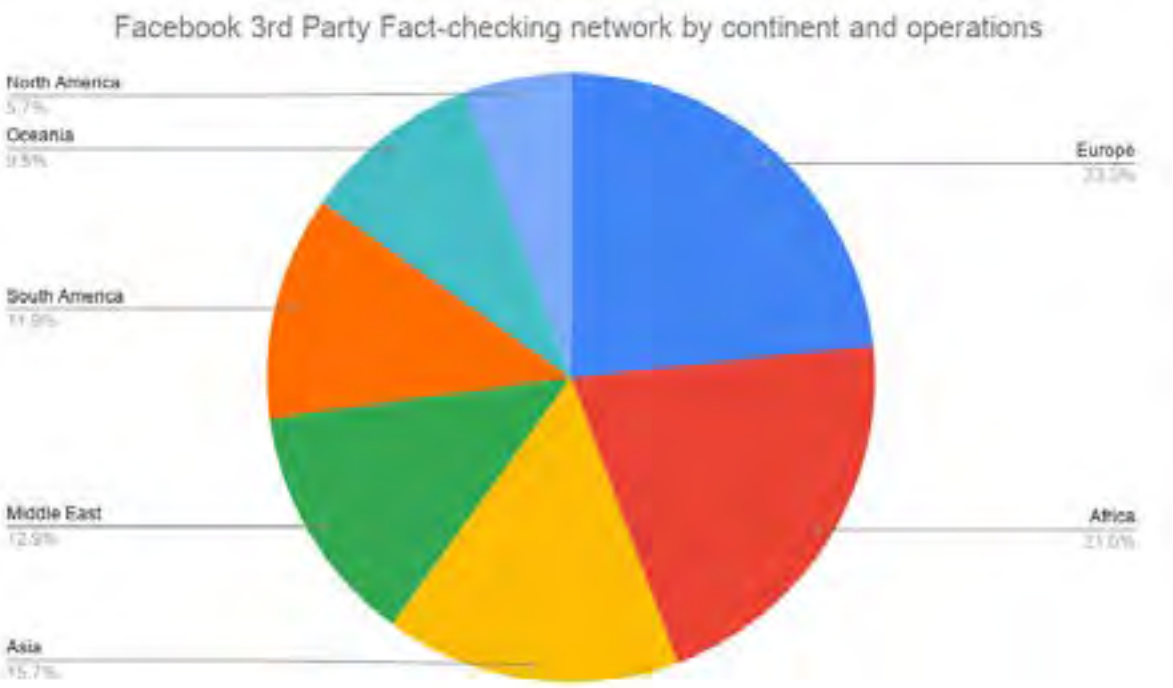


Figure 6. *A view of Facebook third-party fact checking network by continent and operations*

The Third Party Fact-Checking programme was initially launched in the US in mid-December 2016 with The Associated Press, Politifact, FactCheck.org, Snopes and ABC News. It then expanded rapidly internationally. In June 2018, three months after the Cambridge Analytica scandal broke, the Program linked 25 organisations in 14 countries. In early September 2020, when this research was completed, Facebook partnered with 74 third party fact checking organisations around the world, in over 50 languages (this data analysis is based on Facebook's former list of partners and their newest partners' map[45]).

[45] https://web.archive.org/web/20200728165712if_/ https://www.facebook.com/business/help/182222309230722 (deprecated by Facebook in August 2020 and replaced by a map: https://www.facebook.com/journalismproject/programs/third-party-fact-checking/partner-map

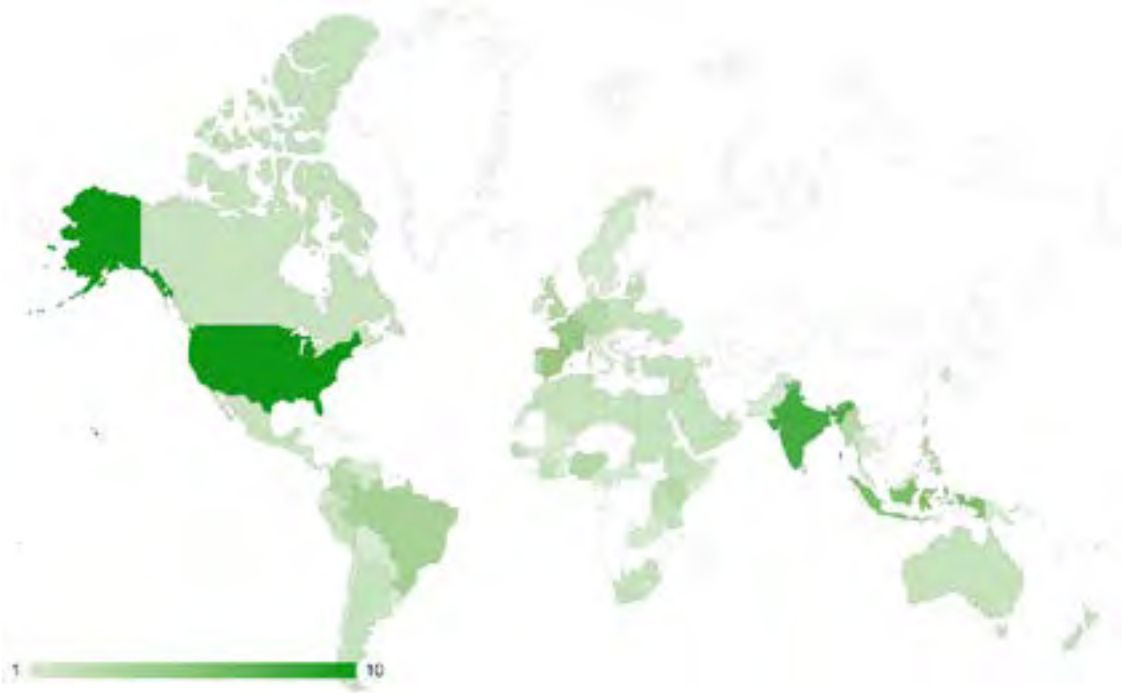


Figure 7. *Map view of Facebook third-party fact checking network worldwide distribution*

The above map (Figure. 7) shows the state of Facebook Third-Party Fact Checking programme in September 2020. Table. 2 below outlines the distribution by number of fact checking operations being contracted by Facebook as of 10 September 2020).

Coverage	Number of operations
United States of America	10
India	8
Indonesia	6
France, Spain	5
Brazil	4
Belgium; Columbia; Democratic Republic of Congo; Kenya; Myanmar; Nigeria; Philippines; Germany; United Kingdom;	3
Algeria; Argentina; Australia; Austria; Bahrain; Canada; Czech Republic; Egypt; Ethiopia; Iraq; Ivory Cost; Jordan; Kuwait; Latvia; Lebanon; Libya; Lithuania; Mexico; Morocco; Netherlands; New Zealand; Oman; Palestine; Peru; Poland; Portugal; Qatar; Saudi Arabia; Senegal; Singapore; South Africa; Sri Lanka; Sudan; Switzerland; Syria; United Republic of Tanzania; Tunisia; Turkey; Uganda; Ukraine; United Arab Emirates; Yemen	2
Azerbaijan; Bangladesh; Benin; Bolivia; Burkina Faso; Burundi; Cameroon; Central African Republic; Chile; Cook Islands; Costa Rica; Croatia; Denmark; Ecuador; El Salvador; Estonia; Fiji; French Polynesia; Ghana; Greece; Guatemala; Guinea-Conakry; Honduras; Ireland; Israel; Italy; Kiribati; Luxembourg; North Macedonia; Malaysia; Mali; Marshall Islands; Micronesia; Montenegro; Nauru; New Caledonia; Nicaragua; Niue; Norway; Pakistan; Palau; Panama; Papua New Guinea; Paraguay; Samoa; Slovakia; Solomon Islands; Somalia; Republic of Korea; Sweden; Thailand; Tonga; Tuvalu; Uruguay; Vanuatu; Venezuela; Zambia	1

Table 2. *Distribution of Facebook's third-party fact checking network by number of operations*

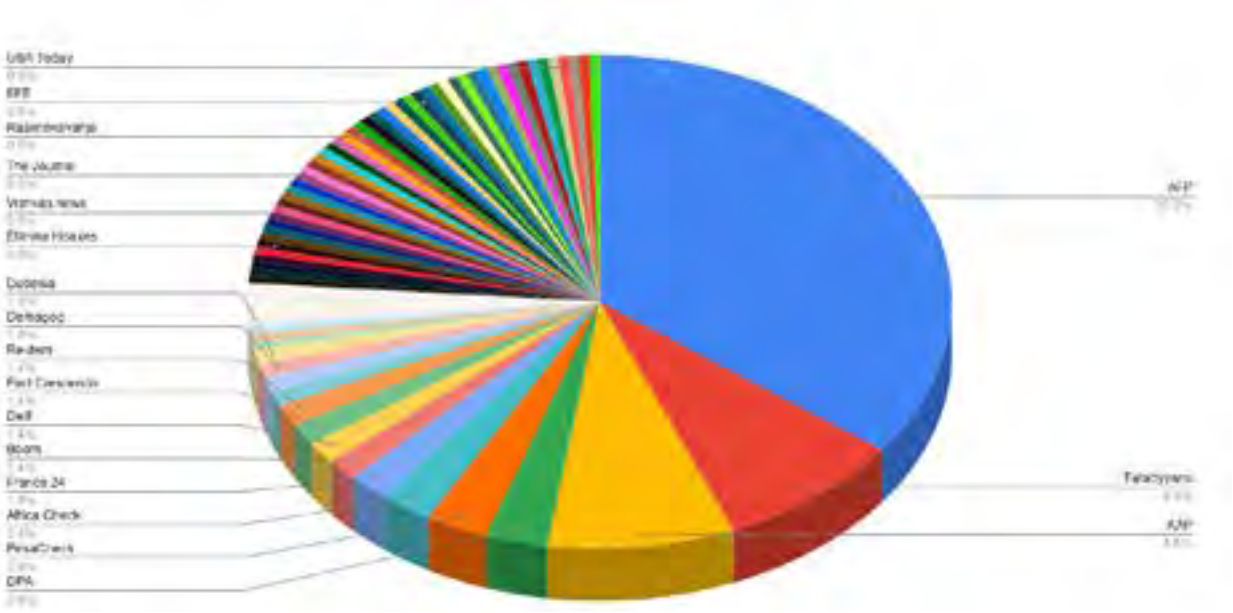


Figure 8. *Distribution of Facebook third-party fact checking programme by organisations involved*

Fact-checkers are selected and remunerated[46] by Facebook. As a prerequisite, all fact-checkers must be certified by the IFCN and adhere to their Code of Principles.

This programme represents the biggest organised international network dealing with disinformation, and it covers both Facebook and Instagram (since May 2019)[47], but it is relevant more widely, as false or misleading content on Facebook is often cross-posted on other social networks like Twitter, YouTube or (Facebook-owned) WhatsApp. So, curtailing the spread of disinformation on Facebook and Instagram can theoretically limit it 'jumping' to WhatsApp and beyond. Although WhatsApp does not directly send contested content to fact-checkers, it has a (little-publicised) chatbot which enables users to get tips on fact-checking and links to local checkers via the IFCN database who can be approached to investigate[48].

In response to COVID-19, IFCN also led the creation of a WhatsApp chatbot that lets users search IFCN's dedicated database of COVID-19 fact-checks (Grau, 2020). In addition, as discussed further in chapter 7.3, in August 2020 WhatsApp started testing (in six countries) a new feature which allows users to carry out simple fact-checking of viral messages themselves, by searching the content on Google (Sweney, 2020). Regarding Instagram, posts rated false by third-party-fact checkers are removed from Instagram's Explore and hashtag pages. In addition, content in Instagram feed and Stories that has been rated false by third-party-fact checkers is down-ranked.[49]

The US in the run-up of the 2020 presidential election has become the top country with ten fact checking organisations, followed by India (with eight), where the spread of

46 See discussion below about transparency issues regarding the fees involved.
47 https://about.fb.com/news/2019/12/combatting-misinformation-on-instagram/
48 https://faq.whatsapp.com/general/ifc-n-fact-checking-organizations-on-whatsapp
49 https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram/

disinformation through word-of-mouth or WhatsApp has fuelled mob violence reportedly resulting in deaths (McLaughlin, 2018), and Indonesia (with six).

The pie chart distribution by organisations in Fig. 8 clearly shows that Agence France-Presse (AFP) news agency has taken a leading share in the programme by launching fact checking operations in more than 70 countries with 90 journalists (including Argentina, Australia, Belgium, Bolivia, Brazil, Burkina Faso, Cameroon, Canada, Chile, Colombia, Costa Rica, Czech Republic, Ivory Coast, Democratic Republic of Congo, Ecuador, El Salvador, France, Ethiopia, Germany, Guatemala, Honduras, India, Indonesia, Kenya, Malaysia, Mexico, Netherlands, Nicaragua, several countries of Middle East and North Africa (mainly from Lebanon), Nigeria, Pakistan, Panama, Peru, Philippines, Poland, Senegal, Singapore, Slovakia, Somalia, South Africa, Spain, Sri Lanka, Switzerland, United Republic of Tanzania, Thailand, Tunisia, Uganda, United States of America, and Uruguay).

AFP made clear in December 2018 that AFP has made the fight against disinformation a core component of its mission, urging that other news agencies have an obligation to debunk false and manipulated stories (Fries, 2018).

Other well known mainstream media embracing fact checking and debunking within Facebook's Third Party Fact-checking programme include: The Associated Press (AP in the United States), The Australian Associated Press (in Oceania), Reuters, German Deutsche Press Agentur (DPA; in Germany, Belgium and Luxembourg), French international broadcaster France 24 Observers team (in 4 countries: Ivory Coast, Democratic Republic of Congo, Guinea-Conakry and France), Rappler (the Philippines), The Quint (India), and French daily *Libération* (France).

Despite the value placed on transparency in fact-checking processes outlined above, there is very limited transparency about how much Facebook pays its third-party fact-checking partners. In a report published in July 2019, British fact-checking operation Full Fact acknowledged that they received £171,800 (for 96 fact-checks) during the first six months of their involvement in the partnership (Hazard Owen, 2019). The money earned depends in part on the amount of fact checking done under the programme. French daily *Libération* and its fact checking service checknews.fr explained that they earned $240,000 in 2018 for 249 articles uploaded to Facebook (Checknews, 2019). Factcheck.org (U.S.) earned an amount of $242,400 during fiscal year 2019 (12 months period ending on June 30, 2019) and $59,500 in the next quarter (1st quarter of fiscal year 2020, ending on September 30, 2019) (Factcheck, 2019). The amount of debunked articles during those periods was not disclosed. Snopes (U.S.), while pulling out from the Facebook partnership in December 2018, disclosed having earned $406,000 in 2018 and $100,000 in 2017 (without reference to the number of debunked suspicious claims). According to fact-checkers' contracts described to the BBC, for each explanatory article, Facebook pays a fixed fee, which, in the U.S., is understood to be around $800 (£600) (Lee, 2019b).

Evaluation of the Facebook initiative

Poynter.org conducted a survey of 19 organisations partnering with Facebook (Funke & Mantzarlis, 2018a) which identified a range of reasons underpinning their decision to join the Network. One of those partners, the British Full Fact referenced above, joined the Facebook Third Party Fact-checking programme late in 2018 and published an evaluative report (Full Fact, 2019) six months into the contract. It remains the most detailed evaluation of the functioning of the programme. While considering Facebook's Third Party Fact-Checking programme as "worthwhile" and likely "needed" for other internet communications companies too, Full Fact also raised some important issues and recommendations like the need for Facebook to:

1. "Fully include Instagram content" into the web interface providing a "queue" of suspicious content provided to fact checkers.[50]
2. "Develop more tools to enable fact checkers to search for and surface similar content" to address "a repeated pattern with online misinformation" and avoid addressing only the tip of the iceberg.
3. Provide more data (to fact checkers) on "shares over time for flagged content"
4. "Share more data with fact checkers about the reach of our fact checks" in order to assess the value of the work undertaken within the program.

Full Fact also regards Facebook's internal rating system - false, mixture, false headline, true, not eligible, satire, opinion, prank generator, and not rated - as 'ill-suited' to the purpose of fact-checking. The first three labels are used by Facebook to reduce the distribution of content and to notify users that this content has been fact checked. Full Fact complained that the 'mixture' label was insufficient as well as over-punitive - it is applied when content is considered a mix of accurate and inaccurate information used for unproven claims and thus the content distribution is downplayed accordingly.

Reacting to Mark Zuckerberg's statement before the U.S. Congress, foreseeing an increasing shift towards a method where more of this content is flagged up front by Facebook A.I. (Artificial Intelligence) tools (Zuckerberg, 2018), Full Fact said it would welcome a clearer statement from the company about "the potential avenues they see for developing machine learning tools" based on the Third Party Fact Checking Partnership data.

Overall, according to the above-mentioned Poynter survey, judged by their own objectives, fact-checkers appear moderately satisfied with the Facebook partnership and the payment they receive for their work. The most critical question for these Facebook partners, according to the Poynter survey, remains concern that the company is not telling the public enough about how the partnership works. But the survey also demonstrated that there is also a lack of certainty about the efficacy of the initiative in terms of actually reducing disinformation on Facebook.

Investigations carried out by Buzzfeed concluded that "Facebook is still the home of viral fake news" (Silverman et al., 2017; Silverman & Pham, 2018). But there was some evidence of a reduction in engagement with disinformation on Facebook after the 2016 U.S. presidential election, with three studies concluding that this could be partially attributed to fact-checking interventions. A study from researchers at NYU and Stanford universities concluded that engagement (shares, likes, comments) with 'fake news' on Facebook fell from a peak of roughly 200 million per month at the end of 2016 to approximately 70 million per month in July 2018 (Allcott et al., 2018). The researchers noted that "...efforts by Facebook following the 2016 election to limit the diffusion of misinformation *may* have had a meaningful impact." The ratio of misinformation and disinformation detected on both Facebook and Twitter also "declined sharply" according to the study, "...from around 45:1 during the U.S. 2016 election to around 15:1 two years later." Nevertheless, according to this research, Facebook remains a much bigger disinformation vector than Twitter.

Another academic study from the University of Michigan introduced an "Iffy Quotient" to describe websites that frequently publish misinformation (Resnick et al., 2019). The study concluded that Facebook and Twitter did a poor job during the 2016 election season,

[50] Since Full Fact's report was published, Instagram content is now subject to fact-checking too, as noted above.

with the distribution of information from questionable sites doubling compared to the rate earlier that year. "However, there has been a long-term decline in Facebook's 'Iffy Quotient' since March 2017", the authors noted.

In further research, Décodeurs, the fact-checking operation of French daily *Le Monde*, analysed 630 French websites in 2018 with the help of their Decodex browser extension which warns web surfers if they reach a dubious news website or a known disinformation source of another kind. They concluded that engagement with low accuracy and dubious websites as well as virality of false news decreased significantly on Facebook (Sénécat, 2018).

Facebook highlighted these studies in a 2018 blog post stating that they represented evidence that the "...overall volume of false news on Facebook is trending downward" (Lyons, 2018b).

More recently, an announcement from Facebook sparked a controversy about the company's policy regarding the fact-checking of political advertising. The company had decided that it would not send organic content or adverts from politicians or their affiliates to its third-party fact-checking partners for review (Clegg, 2019).

Early 2019, a few months after ABC News (U.S.) dropped out of the Facebook fact-checking programme, the anti-hoax U.S. website Snopes decided to quit the Facebook Third Party Fact-Checking programme despite earning 33% of its income in 2018 from the partnership (Green & Mikkelson, 2019). At the end of November 2019, Dutch fact-checker Nu.nl announced their withdrawal from the programme amid controversy around the exemption of certain categories of political advertising (see below) from fact-checking by partners (Hern, 2019a).[51]

Facebook's policy generally exempts political speech from fact-checking, in the form of posts and adverts made by politicians, political parties and affiliates. However, the policy provides that fact-checking can cover "organisations such as Super PACs or advocacy organisations that are unaffiliated with candidates". It also states that:

> *"When a politician shares a specific piece of content - i.e., a link to an article, video or photo created by someone else that has been previously debunked on Facebook - we will demote that content, display a warning and reject its inclusion in ads. This is different from a politician's own claim or statement. If a claim is made directly by a politician on their Page, in an ad or on their website, it is considered direct speech and ineligible for our third party fact checking program — even if the substance of that claim has been debunked elsewhere."*[52]

However, as this study was being finalised in July 2020, Facebook removed a piece of content posted by President Trump for the first time, for violating its COVID-19 related policies. The post included a clip of him claiming that children were "almost immune" to coronavirus. According to a company spokesperson: "This video includes false claims that a group of people is immune from COVID-19 which is a violation of our policies around harmful COVID misinformation." (BBC, 2020d; Carrie Wong 2020)

51 See also chapter 7.1

52 https://www.facebook.com/business/help/182222309230722

Regarding editorial independence, in late 2018 Politifact issued a statement on Twitter endorsed by Factcheck.org (U.S.), Agência Lupa (Brazil) and Teyit.org (Turkey) to deny a report from *The Guardian* claiming that "Facebook pushed reporters to prioritise the debunking of misinformation that affected Facebook advertisers" (PolitiFact, 2018).

According to some news organisations undertaking debunking as members of the programme, Facebook does not prevent them from fact-checking content from politicians and political parties (including political advertising) but they do not pay them to undertake this work and this content is not labelled on the platform when found to be false or misleading by the fact-checkers. For instance, in 2019, AFP fact-checked statements from far-right French leader Marine Le Pen five times on its Factuel blog (Daudin, 2019) and on its Facebook account, but this was not reflected within the Facebook ecosystem due to its policy limiting the fact-checking of much political content.

Opinion content is another contested area. Facebook policy states that opinion is "generally not eligible to be rated by fact-checkers. This includes content that advocates for ideas and draws conclusions based on the interpretation of facts and data, and tells the public what the author or contributor thinks about an event or issue."[53] The policy includes a caveat that "...content presented as opinion but based on underlying false information may still be eligible for a rating." However, this policy has loopholes that have resulted in criticism and controversy. One example is a case pertaining to an op-ed from a climate change denialist group which was based on false and misleading assertions about climate science. In this case, Facebook's climate science fact-checking partner Climate Feedback rated the article as "false"[54], however following an appeal from the lobby group, Facebook removed the label on the basis that the article was an "opinion" and ineligible for fact-checking (Penney, 2020; Pasternak 2020). In another example, a "false" label applied by medical fact-checkers to a video published on an anti-abortion activist's Facebook page claiming that abortion was never medically necessary was removed by Facebook following multiple complaints from conservative lawmakers (Grossman & Schickler, 2019). Although the International Fact Checking Network investigated the fact-checkers' determination and found in September 2019 that the video claim was indeed false[55], the video was still proliferating on Facebook a year later with no fact-checking label[56].

This last example, in particular, prompted lawmakers in the UK House of Lords to note in their report from the inquiry into Digital Technology and the Resurrection of Trust that: "There were no material concerns with the accuracy of the fact check and it was supported by an independent secondary review... This suggests that Facebook's position is more about avoiding political pressure than any particular concern about preserving democratic debate." (House of Lords, 2020).

The Facebook political advertising controversy (concerning its policy on fact-checking noted above) will be covered further in chapter 5.3 on electoral-specific responses, and chapter 7.1, which focuses on ethical and normative responses.

53 https://www.facebook.com/business/help/315131736305613?recommended_by=297022994952764

54 https://climatefeedback.org/evaluation/article-by-michael-shellenberger-mixes-accurate-and-inaccurate-claims-in-support-of-a-misleading-and-overly-simplistic-argumentation-about-climate-change/

55 https://www.poynter.org/fact-checking/2019/the-ifcn-concludes-investigation-about-science-feedback/

56 https://www.facebook.com/youngamericasfoundation/videos/2113086642330235; https://www.facebook.com/youngamericasfoundation/videos/2113086642330235

b. Regional responses

AfricaCheck

AfricaCheck[57] has been the main driver of fact-checking in Africa. It is a non-profit organisation set up in 2012 to promote accuracy in public sphere debate and within the news media in Africa. The goal was to raise the quality of information available to society across the continent. Devised initially by the AFP Foundation, a non-profit media development arm of the international news agency AFP, Africa Check is an independent organisation with offices in Johannesburg (South Africa), Nairobi (Kenya), Lagos (Nigeria) and Dakar (Senegal).

It produces reports in English and French, testing claims made by public figures, institutions and the media against the best available evidence. Since 2012, it has fact-checked more than 1,500 claims on topics from crime and race in South Africa, to population numbers in Nigeria, and fake health 'cures' in various African countries.

Africa Check's work is published and discussed in the news media across the continent. Its head office is based at the Journalism Department of the University of the Witwatersrand in Johannesburg, South Africa, while the French language site has been run by a team based at the EJICOM journalism school in Dakar, Senegal, since 2015. Africa Check relies on its readers to identify the claims they want checked, and it also enables and encourages other journalists to check claims themselves with the assistance of a fact-check section, research reports and teaching services.

Since its creation, Africa Check has received funding support from the AFP Foundation, the Aga Khan University, the Journalism Department of the University of the Witwatersrand, and the EJICOM journalism school, as well as grants from a long list of philanthropic institutions including The African Media Initiative and African News Innovation Challenge, The International Press Institute (IPI), Google, the Konrad Adenauer Stiftung, the Millennium Trust, Luminate, the Open Society Foundations (OSF), the Shuttleworth Foundation, the Bill and Melinda Gates Foundation, the Raith Foundation, Standard Bank, Absa and Code for Africa.

The Duke University fact-checking database registered 17 active fact-checking organisations in Africa in early 2020.

Latin America: the influence of Chequeado

In Latin America, Argentina's Chequeado[58] has been prominent in the fact-checking community since its creation in 2010. Many new initiatives have emerged in the region since 2014, mostly in the journalism field, thanks to the help and influence of Chequeado, especially in the area of fact-checking methodologies. In 2019, Chequeado coordinated with AFP on the Reverso project to fact-check the Argentinian presidential election campaign.

Duke University's fact-checking database registers 16 organisations in South America, eight in Central America in Spanish, and 10 in Portuguese (in Europe, there are 6 organisations in Spain and 2 in Portugal). In 2014, Chequeado invited the other regional fact-checking organisations to a meeting in Buenos Aires to launch a new network "LatamChequea" designed to exchange best practices. Since then, the regional network

57 https://africacheck.org/
58 https://chequeado.com/

has been holding a biannual conference in Buenos Aires. There are also monthly virtual meetings between fact-checkers which also involve a number of social scientists.

Chequeado is supported financially by a foundation, La Voz Pública, and it is active in research collaborations with academics. Fact-checkers are also embedded for a week or two in Chequado's newsroom with the support of IFCN scholarships.

Europe: SOMA

The Social Observatory for Disinformation and Social Media Analysis (SOMA) is funded by the European Commission with the objective of organising European fact-checkers as part of a pan-European effort to rebuild trust in journalism, and to provide support to the growing community of media organisations, fact-checkers, academics, and NGOs and policy makers fighting disinformation on the continent. In the first year of operation, some 40 European organisations have formally join this Observatory, based on the platform Truly Media[59]. This European Observatory has published several investigations and recommendations regarding disinformation around the COVID-19 pandemic. This observatory is due to be continued in the forthcoming years by a new one called EDMO (European Digital Media Observatory).[60]

Arab States

In the Arab countries, collaboration between fact-checking initiatives is not institutionalised but fact-checkers in the region are connected, collaborate on training, and gather at conferences such as Alexandria Media Forum in Egypt which has focused on fact-checking, disinformation, and media literacy and training in its three last editions in Alexandria (2017-2018) and Cairo (2019).

Regionally, one prominent initiative, launched in 2014, is Jordan-based Fatabyyno[61]. Launched in 2014, it monitors and debunks disinformation in eighteen countries in the Middle East and North Africa. Others include Da Begad[62] ('Is it real?') launched in Egypt in 2013, as well as Matsad2sh[63] ('Don't believe') and Falsoo[64]. Homonyms Falso work on fact-checking in Libya[65] and in Tunisia[66].

In the Syrian Arab Republic, Verify Syria[67] is publishing a monitoring and debunking website in three languages, Arabic, English and Turkish. The AFP fact-checking operation[68] covering various countries of Middle East and North Africa is based in Lebanon as a collaboration with the local fact-checker El3asas[69].

c. Some other national responses

This subsection details specific and noteworthy national initiatives in the area of monitoring and fact-checking. In the U.S. and in Europe, the history and evolution of fact-

59 https://www.truly.media/
60 https://edmo.eu/
61 https://fatabyyano.net/ and https://www.facebook.com/Fatabyyano/
62 https://dabegad.com/
63 https://www.facebook.com/matsda2sh/
64 https://www.falsoo.com/
65 https://falso.ly/
66 https://www.facebook.com/falso.tn/
67 https://www.verify-sy.com/
68 https://factual.afp.com/ar
69 https://twitter.com/el3asas

checking is tied to election campaigns and verifying political claims. Therefore most of these responses are analysed in chapter 5.3.

India

In India, Facebook-owned WhatsApp has developed into one of the main channels of disinformation (Kajimoto & Stanley, 2019). It has been noted that the phenomenon of politicians using social media to directly access audiences, bypassing traditional media gatekeepers, has been identifiable since 2014 and it has aided the spread of disinformation within online social networks (Kaur & Nair, 2018).

Fifteen active fact-checking organisations operate in India according to Duke University's database and eight are members of Facebook's third-party fact-checking network. Nevertheless, those outlets are mostly individuals, and small organisations or teams (like the Times of India fact-checkers). All of these have been created since 2013. They include Factcheck.in[70], SM Hoax Slayer[71], and investigative journalism outlet Boomlive, which pivoted to fact-checking in 2016. A large part of the disinformation they debunk is political, either local or about geopolitical tensions.

Indonesia

In Indonesia, the NGO Mafindo has been fighting disinformation since 2015 through an anti-defamation and hoax group[72] on Facebook, a WhatsApp-based hoax buster[73], a Google Chrome extension, and a website[74] using their motto "Turn Back Hoax". Following Mafindo, five other debunking initiatives have been launched in Indonesia, mostly by news organisations. Six of them are part of the Facebook third-party fact-checking network. Another initiative mentioned by researchers (Kajimoto & Stanley, 2019), Klarifikasihoax has been inactive since 2017.

Philippines

In the Philippines, disinformation campaigns are collectively creating a public sphere filled with information pollution and, consequently, with toxic incivility and polarisation since the 2016 presidential election. As reported in a UNESCO publication (Posetti 2017), troll armies using 'sock puppet' networks have gained traction with potentially long-term consequences for democracy and elections (see also Ong & Cabañes, 2018; Ressa, 2016). However, four fact-checking organisations are monitoring disinformation and political claims, including Vera files[75] and Rappler[76], and three of them are members of the Facebook third-party fact-checking network.

Republic of Korea

In the Republic of Korea, there has been a proliferation or rumours, partisan propaganda and disinformation on mobile messaging apps like KakaoTalk or Naver Band, as well as social media sites, especially during elections. One of the main initiatives set up for the 2017 presidential election was SNU Factcheck[77], launched by Seoul National University to gather 26 news outlets to cross-check disputed information. It continues as one of the

70 https://www.factchecker.in/about-us/
71 https://smhoaxslayer.com/about/
72 https://www.facebook.com/groups/fafhh
73 https://mafindo.gitbook.io/whatsapp-hoax-buster/
74 https://turnbackhoax.id/
75 https://verafiles.org/
76 https://www.rappler.com/newsbreak/fact-check
77 http://factcheck.snu.ac.kr/

five fact-checking organisations in the country. The Facebook third-party fact-checking network does not have a Korean member, mainly because the local Naver and Daum are the most popular online portals with a policy of asking news content providers to go through an evaluation process and thereby making it harder for disinformation purveyors to syndicate content through those portals.

U.S. - Snopes

Snopes is one of the debunking and fact-checking pioneers in the U.S.. Back in 1994, founder David Mikkelson created snopes.com[78] as a hobby to investigate urban legends and hoaxes on the Usenet (a worldwide discussion channel of the early internet) newsgroup alt.folklore.urban.

Immediately after the 9/11 terrorist attacks in 2001, the founders of snopes.com started to debunk rumours and lies about the attacks - a total of 176 legends and rumours[79] were evaluated by Snopes between 2001 and 2011 (Aspray & Cortada, 2019). This was the inflection point for snopes.com to shift from demystifying urban legends as a hobbyist, to progressively becoming a major fact-checking organisation (Dean, 2017). Between breaking news (like Hurricane Katrina in 2005) and presidential elections (2008 with the rumours circulating about Barack Obama's place of birth; up to the 2016 poll), snopes.com grew its audience, allowing it to build a sustainable business through advertising revenue.[80]

4.1.5 Response Case Study: COVID-19 Disinformation

WHO Director-General Tedros Adhanom Ghebreyesus has stated that the Coronavirus outbreak has left humanity not just fighting an epidemic but also "an infodemic" (Zarocostas, 2020). All organisations reviewed in this chapter have taken measures to respond to the COVID-19 crisis with special hubs or pages about COVID-19 disinformation.

For example, First Draft has published a whole hub[81] of resources for reporters such as tools, guides, ethics guidelines, an online course, and a searchable database of coronavirus debunks based on two monitoring tools: Google Fact Check Explorer[82], and the IFCN CoronaVirusFacts Alliance database[83]. The latter was launched in January 2020 as a double hashtag campaign on Twitter #CoronaVirusFacts (in English) and #DatosCoronaVirus (in Spanish) for participating IFCN members, when the epidemic was still limited to China but was already being exploited for disinformation purposes.

The hashtags campaign led to a database of more than 3000 fact-checks from 70 countries and 40 languages (in April 2020). Then, another project led by Science Feedback[84], and sponsored by the Google News Initiative, sought to expand this database with all the urls sharing COVID-19 disinformation.

78 https://www.snopes.com/
79 https://scholar.colorado.edu/concern/book_chapters/8049g572m
80 See the earlier discussion in this chapter, and in chapter 7.1, Snopes' role as a member of the Facebook Third Party Fact-checking Network
81 https://firstdraftnews.org/long-form-article/coronavirus-resources-for-reporters/
82 https://toolbox.google.com/factcheck/explorer
83 https://www.poynter.org/ifcn-covid-19-misinformation/
84 https://sciencefeedback.co/building-an-open-source-database-of-misinformation-sources-on-covid-19/

The fact-checking community, from the IFCN, Facebook's Third Party Fact-Checking programme and beyond, have published countless debunking reports about the Coronavirus outbreak, registering disinformation cases from all continents. According to a study from the Reuters Institute for the Study of Journalism (RISJ) based upon an analysis of English-language fact-checks curated by First Draft, the number of fact-checks increased more than 900% from January to March 2020. On the 225 debunks analysed, RISJ found that 59% of the misinformation content was reconfigured while 38% was fabricated. (Brennan et al 2020)

Some internet communications companies (e.g. Facebook[85], YouTube[86], Instagram[87], WhatsApp[88], Twitter[89], Linkedin[90]) themselves have taken action to connect their users to reliable information about the pandemic by linking any query on the coronavirus to the World Health Organisation (WHO) main hub[91] and their WHO mythbusters page[92], or to the local government's ministry of health. They are also relaying alerts from WHO through chatbots, and from local authorities on message applications[93] too, or publishing curated official pages of factual information. Some are also promoting the IFCN affiliated fact-checking organisations[94] and asking their users to verify the facts and to refrain from sharing information if they are not sure it is true.

Google (Mantzarlis, 2020), Facebook (Goldshlager & Watson, 2020) and WhatsApp (IFCN, 2020a) announced small programmes to fund fact-checkers and nonprofits fighting disinformation about the pandemic in several countries (IFCN, 2020b). Thirteen projects in the same number of countries were announced at the beginning of April through the "Coronavirus Fact-Checking Grants" (IFCN, 2020c) program.

In addition, internet communications companies have decided to "work closely together"[95] to combat fraud and misinformation connected to the pandemic. Many companies have started blocking adverts that try to capitalise on coronavirus-related disinformation and removing disinformation that could lead to physical harm. For example, in April Facebook said that it put 50 million warning labels on pieces of content on the platform, based on over 7,500 articles from their fact-checking partners.[96] Some are also removing conspiracy-type content, using policy provisions about content consideration in terms of its likely potential to cause harm.[97] In Facebook's case, this provision is ""Misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm."[98] Facebook CEO Zuckerberg stated that it was "easier" to make the difference between good and wrong information in a pandemic than in a political campaign (Smith, 2020a).

It is not possible to accurately gauge the extent of fact-checked COVID-19 disinformation within the companies, because they typically do not provide access to granular statistics

85 https://www.facebook.com/coronavirus_info/?page_source=coronavirus_hub_attachment&fref=mentions
86 https://www.youtube.com/watch?v=i352PxWf_3M
87 https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram/
88 https://www.whatsapp.com/coronavirus
89 https://blog.twitter.com/en_us/topics/company/2020/covid-19.html
90 https://www.linkedin.com/feed/news/coronavirus-official-updates-4513283/
91 https://www.who.int/emergencies/diseases/novel-coronavirus-2019
92 https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters
93 https://www.messenger.com/coronavirus
94 https://faq.whatsapp.com/126787958113983
95 https://twitter.com/googlepubpolicy/status/1239706347769389056
96 https://about.fb.com/news/2020/04/covid-19-misinfo-update/
97 https://blog.twitter.com/en_us/topics/company/2020/covid-19.html
98 https://www.facebook.com/communitystandards/credible_violence/

about the origins, numbers and types of items checked, nor data on the circulation of such content prior to and post any actions being taken. (UNESCO, 2020)

Notwithstanding fact-checking efforts, briefing papers issued by the Institute for Strategic Dialogue (ISD, 2020a) in March and April 2020 have warned about the exploitation of COVID-19 pandemic by anti-migrants and xenophobic or far-right networks (ISD, 2020b), especially in closed groups on Facebook, chat channels on WhatsApp, and fringe networks, like 4chan (Arthur, 2019), as well as in languages other than English.

4.1.6 How are monitoring and fact-checking responses evaluated?

Fact-checking can be evaluated in terms of whether it is achieving its immediate and longer term objectives. This depends on assessing its volume, reach and timeliness, among other factors. However, this is not straightforward as there is still limited published research on the reach and impact of fact-checking. Much of the relevant data is held in private by the internet companies. This makes evaluation difficult, and leaves researchers to make extrapolations from limited data (such as reach and engagement metrics attached to debunks and fact-checks published by news organisations), and audience research (e.g. ethnographic, psychological studies) into diverse citizens' responses to both disinformation and corrective measures. Further, the underlying assumption that verified evidence and rational thought have a role to play in countering disinformation is hard to test empirically because of the complex interlinkages of disinformation with emotion and identity.

During the run-up to the 2012 U.S. presidential election, concerns arose about the efficiency of fact-checking and its ability to reduce disinformation, particularly that connected to political rhetoric. But there was still sufficient momentum for the continuation of fact-checking efforts, as evident in the words of one researcher: "Naming and shaming politicians who repeatedly mislead the public can still inflict significant reputational damage over time" (Nyhan, 2012).

Promoting more involvement of citizens in public affairs, increasing politicians' reputational cost, and increasing the public's trust in the news media have been identified in several studies as having positive effects for fact-checking. However, the predisposition of citizens to accept corrections that reinforce their own views is relevant. The backfire or 'boomerang effect' helps to spread disinformation (i.e. if fact-checks contradict citizens' pre-existing views about a political actor or issue, they are more likely to be rejected despite their accuracy), especially when disinformation (and fact-checks) are weaponised by the politicians themselves to increase polarisation.

One fact-checking organisation that has tried to assess the impact of its work is Argentina's Chequeado. In a review of six academic studies assessing the impact of fact-checking in the United States, researchers commissioned by Chequeado to study the efficacy of their efforts considered the impact of fact-checking on citizenry, political elites, and media organisations. (Pomares & Guzman, 2015) They found that promoting the involvement of citizens in public affairs, increasing the reputational cost of spreading falsehoods for politicians, and aiding public trust in the news media are positive effects of fact-checking. One evaluative response proposed by the Chequeado-commissioned researchers is to measure the strength of the fact-check (e.g. how well did it stand up to scrutiny?) and its reach in tandem. (Pomares & Guzman, 2015)

A recent joint report from Africa Check, Chequeado and Full Fact listed public corrections of misleading statements or statistics, stopping false claims from politicians, releasing new meaningful data, getting journalists in legacy newsrooms trained to reduce the spread of disinformation, and engaging with officials, and efforts to raise accountability as potential benefits of fact-checking (Africa Check, Chequeado & Full Fact, 2020). These could be treated as indicators for efficacy.

4.1.7 Challenges and opportunities

The volume and range of types of disinformation make it difficult to identify, monitor, report and draw public attention to all instances and all dimensions of the problem. There are also key nuances, such as it is one thing to demonstrate that a claim is false, another to show that it is currently without evidence (but potentially could be true), and a third to say that whether a particular proposition is factual when there is usually a wider narrative or perspective at play which mobilises and combines particular facts, as well as presents them along with opinion, attitude and identity.

This is further complicated by the task of assessment of the intended and unintended effects of identification of the content at hand, and of its providers. However, producing such analysis is vital in order to develop or modify fact-checking and other responses.

The challenge for fact-checkers is to aspire to objective standards and operate transparently in all countries and languages, at scale, and with impact. This is necessary to enable society to access the information required to ensure that the various responses are optimally effective. Achieving this in practice, however, is far from straightforward, especially in the case of non-global languages and smaller countries, which often do not have their own local independent fact-checking organisations. Instead, international fact-checking organizations aim to fill the gap, but inevitably need to rely on native speakers, limiting the possibility of scrutinising their work and biases. This highlights the need for a robust, independent approach to 'evaluating the evaluators' or 'fact-checking the fact-checkers.'

Fact-checking also needs to be consistent with international standards for freedom of expression and other human rights like privacy, and to recognise that certain content (e.g. unknowns, certain narratives, opinions, humour) does not lend itself to verification per se. Further, fact-checking has to live up to values of transparency and non-partisanship, and avoid selective instrumentalisation.

A challenge for fact-checking organisations is to fend off legal attacks on them. The Fact-checkers Legal Support Initiative has come into existence to assist with legal advice[99]. It is a consortium of the Media Legal Defence Initiative, the International Fact-Checking Network and the Reporters Committee for Freedom of the Press.

Major events, such as elections and public health emergencies, provide an opportunity for independent monitoring and identification responses to reaffirm the value of facts, and to encourage public reflection of what content they treat as credible, and what people decide to share. For example, identifying COVID-19 disinformation and investigation of responses over time also enables continuous assessment of the internet communications companies' efficacy in "flattening the curve" of the 'disinfodemic' (Proctor, 2020; Posetti & Bontcheva 2020a; Posetti & Bontcheva 2020b). Identification responses are also key

[99] https://factcheckerlegalsupport.org/

for monitoring the intersection of disinformation with hate speech used against women, minorities, migrants and other vulnerable citizens and communities. However, it is acknowledged that some of these targeted groups may also resort to disinformation tactics and content produced by them should also be scrutinised. It is the case that sometimes groups that are victims of disinformation may themselves resort to the same tactics to further their various causes, and that their content should also be subject to fact-checking and scrutiny.

This is also an opportunity to strengthen identification responses. While WhatsApp (IFCN, 2020a), Facebook (Axelrod, 2020), Google[100], and Twitter (Gadde, 2020) have pledged some funding to fact-checking organisations, this also shows that more can be done. Ongoing support throughout and beyond critical periods of elections and pandemics is needed. Verifying claims about vaccinations and climate change is particularly significant going ahead.

4.1.8 Recommendations for monitoring and fact-checking responses

The challenges and opportunities identified above, and the current state of fact-checking and debunking, lead to the following policy recommendations for international and regional institutions, governments, internet communications companies, foundations and news organisations, which could:

- Make available resources for independent fact checking, including facilitating the fact-checking of political content and political advertising.
- Support the principle of access to information, especially in regard to both authorities and internet communications companies, as relevant to fact-checking, in order to increase transparency and enable fact-checking organisations themselves to work more accurately and transparently.
- Promote fact-checking results as trustworthy sources of information, useful for citizenship, for the news media, and for Media and Information Literacy interventions.
- Promote trans-disciplinary research into fact-checking responses to disinformation.
- Help to develop collaborative fact-checking operations worldwide to aid access to accurate and reliable information globally, especially in partnership with news organisations.
- Reinforce fact-checking capacity within news organisations through specialist training and editorial projects to support accountability reporting applied to corporate, government, and political actors and actions.
- 'Verify the verifiers' and develop international standards and an accountability approach to enable transparent, and objective appointment and assessment procedures for the people and organisations (including the internet communications companies that facilitate and fund fact-checking on their sites) involved in fact checking, and evaluate their performance over time.

[100] https://www.france24.com/en/20200402-google-boosts-support-for-checking-coronavirus-facts

4.2 Investigative responses

Authors: Sam Gregory, Julie Posetti and Denis Teyssou

This chapter addresses the range of entities producing investigations into disinformation (ranging from journalism to civil society investigations) and their outputs.

Investigative reports typically address particular campaigns and actors, and go beyond the questions of whether particular content contains falsehoods and the extent of the falsity (fact-checking). They may include, but also go beyond, the issue of whether a piece of content is manipulated or miscontextualised/misrepresented (such as in the case of provenance labelling), or whether a piece of content or outlet is reliable and fair (for example, credibility labelling). They are likely to monitor (as far as possible) the instances, themes and spread of the particular disinformation under focus. When fully deployed, they also provide insights into the dynamics of disinformation campaigns, including such elements as the networks conducting them, the targets, the mediums used, the methods used, budgets available, along with attribution and intent. For examples of categorisations and cataloguing of campaigns, see Bradshaw & Howard (2018) and Brooking et al. (2020).

Such investigations typically aim to help news organisations, governments, fact-checkers, Internet communications companies and others understand these dynamics, in order to deploy effective counter-measures to particular campaigns. They seek to alert actors to ongoing innovations in disinformation tactics and strategies. Increasingly, methodologies of categorisation are being developed to better catalogue across related incidents.[101]

4.2.1 What and who do they target?

Investigative responses monitor a range of actors. A significant number focus on government-funded or -supported disinformation campaigns. Although many commercial and company responses initially began with a focus on one or two governments' roles in targeted disinformation campaigns, the range of state sponsors has expanded (Nimmo, 2019; Francois et al., 2019; Nimmo et al., 2019a; Gleicher, 2019a; Gleicher, 2019b). The '2019 Global Inventory of Organised Social Media Manipulation' report (Bradshaw & Howard, 2019) identifies government-implicated social media manipulation campaigns against foreign countries conducted by a number of States, while noting over 26 countries with internal disinformation activities. However, the entities above tend to have a blind-spot in regard to the covert or overt disinformational activities by certain governments in foreign countries or domestically. There is a focus on disinformation promoted by unofficial actors such as on white supremacist groups in the U.S. or far-right movements in India (the Southern Poverty Law Center, Equality Labs -see Soundararajan et al., 2019).

101 https://www.bellingcat.com/

Other investigators look at commercially-motivated or hybrid actors (albeit often working for political purposes). For example, foreign 'click-farms' engaged in significant disinformation interventions in U.S. politics for commercial reasons (Silverman & Alexander, 2016; Soares, 2017). Another example is investigation into the Epoch Media Group, a commercial entity with political motivations, which led a disinformation campaign including fake profiles and accounts (Nimmo, et al., 2019b). This was exposed via reporting from Snopes and the Operation #FFS investigation by Graphika[102]/ Facebook. An important trend in the past 2-3 years has been the growth of private sector disinformation-for-hire actors, providing services to politicians and political parties - as documented in a number of countries (Ong & Cabañes, 2019; Silverman et al., 2020). A 2019 survey by Buzzfeed News based on account takedowns by platforms, as well as publicised investigations by security and research firms "found that since 2011, at least 27 online information operations[103] have been partially or wholly attributed to PR or marketing firms. Of those, 19 occurred in 2019 alone." (Silverman, et al., 2020). Another important investigation was undertaken by South Africa's Daily Maverick into the now defunct UK-based PR firm Bell Pottinger which was exposed for artificially seeding racial tension in the country amid a state capture scandal linked to the presidency (Thamm 2019; Posetti et al 2019a).

4.2.2 Who do investigative responses try to help?

Investigative reporting serves a range of actors including companies engaged in detection of coordinated inauthentic behaviour on their platforms as well as official inquiries. One such inquiry was the U.S. Congressional investigation into foreign interference before and during the 2016 U.S. elections (U.S. Senate Select Committee on Intelligence, 2018), and another was the UK-initiated International Grand Committee on Disinformation and 'Fake News'[104]. Governments also use these investigations, for example EU policy proposals based on commissioned reports (EU Disinfo Lab, 2019b). Coordination between internet communications companies and external actors investigating campaigns is often reflected in funding relationships between them (see below and chapters 4.1, 7.1).

Civil society organisations like Amnesty International also undertake forensic investigative work designed to detect, debunk and deter disinformation connected to human rights abuses. These include Amnesty's Digital Verification Corps[105] - a partnership with six international universities that also collaborates on open source journalistic investigations (Fortune, 2018). Other stakeholders include individual citizens and the growing number of participants in the global anti-disinformation community. One such example is Amnesty International's 'Amnesty Decoders'[106] project, which crowdsources volunteer verification assistance to examine claims of human rights violations. Campaigning organisation Avaaz has also done investigations, including into the responses by internet communications companies, and advocated for changes accordingly.[107]

102 https://graphika.com/
103 https://www.militaryfactory.com/dictionary/military-terms-defined.asp?term_id=2637
104 https://www.oireachtas.ie/en/press-centre/press-releases/20191025-international-grand-committee-on-disinformation-and-fake-news-dublin-ireland-wednesday-6th-and-thursday-7th-november-2019/
105 https://www.theengineroom.org/digital-verification-corps/
106 https://decoders.amnesty.org/
107 https://secure.avaaz.org/campaign/en/disinfo_hub/

The motivation behind investigative responses is to improve understanding of specific disinformation acts and campaigns so as to enable action to be taken against them. Such action could include content takedowns or demotion, legal processes, transparency and accountability measures, and regulatory or company policy reform. Investigative responses also aim to expose methods adopted in disinformation campaigns to impede further utilisation of these by malicious actors, and ensure knowledge on them is available to a burgeoning community of disinformation researchers. They identify structural challenges in disinformation as opposed to symptomatic examples or individual content items. As an example, the EU Disinfo Lab initiative notes its commitment to "continuously monitor disinformation activities across the major platforms (digital and traditional), identify trends and threats, alert activists and researchers to these, and make our resource pool widely available to collaborators."[108]

Within this broad framework, organisations producing investigative reports are positioned differently in relation to the State. There is contrast between a specialist news publisher like Bellingcat[109] which focuses on publicly available data and open-source investigation as an entry point for establishing facts, and investigative entities that work closely with particular state agendas and/or are aligned with particular companies.

4.2.3 What output do investigative responses publish?

Investigative reports take a range of forms. Most NGO and journalistic investigations focus on providing information on the mechanics and sources of disinformation (to expose approaches to creating and distributing disinformation) and creating in-depth reporting for broad consumption. Transparency on methods (and indeed co-creation and participation via crowd sourcing in evidence-gathering) is a key part of OSINT (Open Source Intelligence) approaches, as practiced by BellingCat and other hybrid organisations, while some entities produce data based on their specialities (for example, social graph network analysis methods in the case of Graphika). Investigations by internal groups within internet communications companies do not typically provide complete data on how they identified disinformation. It is an ongoing critique of the companies' approaches to their own identification of disinformation, as well as their support to others, that there is a lack of provision of data to help identify, categorise and define disinformation campaigns.

However, a growing number of groups are trying to establish shared methodologies for classification. One example is the work of the Digital Forensic Research Lab to develop a framework for categorisation in their Dichotomies of Disinformation project (Brooking, et al., 2020) (with support from Jigsaw, a division of Alphabet, the holding company of Google). This categorisation approach includes over 150 binary, text-based and quantitative variables grouped under a top-line set of variables that includes: target, platform, content, method, attribution and intent. Other categorisation work includes the Computational Propaganda Project's surveys of organised social media manipulation based on an assessment of news media reporting (Bradshaw & Howard, 2019), and the Institute for the Future's reporting on types of state-sponsored trolling within disinformation and online harassment campaigns (Monaco & Nyst, 2018).

[108] https://www.disinfo.eu/
[109] https://www.bellingcat.com/

4.2.4 Who are the primary actors and who funds investigative responses?

A range of initiatives work on organised investigations into disinformation and produce in-depth reporting. These include:

Entities with a primary focus on disinformation: The Digital Forensic Research Lab of the Atlantic Council is an example of an organisation with a strong focus on identifying, unpacking and countering disinformation campaigns. It publishes reports, develops field expertise and identifies methodologies and tracking approaches (Brooking, et al., 2020). In the European context, EU Disinfo Lab is a more recently established NGO designed to maintain both a platform for analysis of disinformation in Europe, although it also monitors and identifies disinformation operations with international dimensions (EU Disinfo Lab, 2019a; Carmichael & Hussain, 2019). Actors responding to disinformation in this subcategory span foundation and government-funded outfits, non-governmental organisations, and dedicated academic programmes. Some investigations have delved into the business models used by many internet companies, attributing to these a propensity towards the propagation of rumour and conspiracy theorists. For example, the Center for Humane Technology says that YouTube recommended conspiracy videos by Alex Jones more than 15 billion times.[110]

Entities with methodologies relevant to disinformation, such as Open-Source Intelligence (OSINT): In parallel with the development of the disinformation research and investigation field in the past six years, there has also been the growth of an increasingly robust field of open-source investigation more broadly, using 'open source' and social media sources to conduct investigations into topics such as war crimes and chemical weapons usage. An example of an organisation in this field is Bellingcat, self-described as an "...independent international collective of researchers, investigators and citizen journalists using open source and social media investigation to probe a variety of subjects – from (...) drug lords and crimes against humanity, to tracking the use of chemical weapons and conflicts worldwide."[111] It has staff and contributors in more than 20 countries around the world, who work at the intersection of advanced technology, forensic research, journalism, investigations, transparency and accountability monitoring. Entities in this group include foundation and government-funded outfits, and NGOs.

Investigations by existing non-governmental watchdogs or monitors with a thematic or sectoral Freedom of Expression focus: Although disinformation should not be conflated with hate speech, the combination of the two involves a range of existing groups who investigate patterns of malicious information-sharing in particular thematic contexts. One example is the Southern Poverty Law Center in the U.S., which exists to "monitor hate groups and other extremists throughout the United States and expose their activities to the public, the media and law enforcement".[112] They provide a comprehensive biannual report into the status of these movements and their activities, as well as specific reports into particular propaganda activities. Similarly, the London-based Institute for Strategic Dialogue (ISD)[113] documents and produces reports on extremist violence and related speech. (As chapter 7.1 outlines, there are significant overlaps between normative and ethical responses to disinformation, and the issue of hate speech).

110 http://humanetech.com/wp-content/uploads/2019/07/CHT-Undivided-Attention-Podcast-Ep.4-Down-the-Rabbit-Hole.pdf ; https://www.newamerica.org/oti/events/online-getting-to-the-source-of-the-2020-infodemic-its-the-business-model/

111 https://www.bellingcat.com/about/

112 https://www.splcenter.org/fighting-hate

113 https://www.isdglobal.org/isdapproach/

In-depth investigations by news outlets: A range of news outlets maintain ongoing disinformation investigatory beats. One example is Buzzfeed News, providing insights and investigations into individual disinformation campaigns and trends in disinformation, such as the growing use of pay-for-hire PR firms in disinformation (Silverman, et al., 2020). Other outlets have conducted in-depth investigations of particular campaigns, such as Rappler's mapping of disinformation networks during and after the 2016 presidential elections in the Philippines (Ressa 2016; Posetti et al., 2019a), the work of South Africa's Daily Maverick referenced above (Thamm 2019; Posetti et al., 2019a), and that produced by the African Network of Centers for Investigative Reporting on media manipulation in South Africa[114] (ANCIR, n/d). Another contribution comes from CodaStory, which has a disinformation specialisation and focuses on investigations into orchestrated campaigns connected to state actors and disinformation agents for hire (Dorroh 2020).[115]

Action-oriented academic research: A burgeoning number of academic departments produce both meta-analyses of disinformation campaign strategies, for example the inventories of organised social media manipulation (based on news media content analysis) from the Computational Propaganda Project at the Oxford Internet Institute (Bradshaw & Howard, 2019), as well as detailed research into specific strategies and country-contexts. An example of the latter is academic work on networked social media manipulation in the Philippines (Ong & Cabañes, 2018). Other research from the Reuters Institute for the Study of Journalism at the University of Oxford focuses on the exposure of the public to disinformation in a number of countries, as well as investigative responses of journalism (Posetti et al., 2019a) and media effects (Nielsen & Graves, 2017).

Commercial entities working in social network analysis and cyber-security: A range of commercial companies provide services or conduct investigative research into disinformation campaigns. An example is Graphika, which focuses on detecting "strategic influence campaigns online and at scale by analyzing network anomalies and identifying objects propagating through network maps with a high degree of social contagion that are likely to quickly reach virality."[116] The company applies social media network analysis to conduct investigations into specific campaigns. These investigations can be in coordination with other actors - for example, with companies such as Facebook, in the 'Operation #FFS: Fake Face Swarm' (Nimmo et al., 2019b), an analysis of fake profiles/accounts tied to the Epoch Media group. Another example of a commercial entity is FireEye, which has a commercial cybersecurity background. It has identified and investigated cybersecurity breaches and related disinformation campaigns originating in various States (Revelli & Foster, 2020).

Investigations by internal company threat mitigation teams: All major social media companies have internal threat analysis teams, and teams dedicated to 'site integrity' or identifying 'coordinated inauthentic behavior' (Gleicher, 2018a). For example, Facebook has produced a report on tackling co-ordinated inauthentic behaviour in a number of countries (Gleicher, 2020). These teams sometimes share specific data to outside partners or collaborate/contract with external companies and non-profit/academic groups (Gleicher, 2018b). In the case of Facebook this includes collaborations with a number

114 https://s3-eu-west-1.amazonaws.com/s3.sourceafrica.net/documents/118115/Manufacturing-Divides.pdf

115 See also this video panel discussion about in depth journalistic investigations into disinformation in the context of COVID-19 featuring CodaStory Editor Natalia Anteleva, Buzzfeed's disinformation specialist Jane Lytvynenko, and Rappler's Executive Editor, Maria Ressa: https://www.youtube.com/watch?v=tBp4OKSW_ho&feature=youtu.be

116 https://www.graphika.com/graphika-labs

of the other types of entities cited in this chapter - e.g. Digital Forensic Research Lab, Graphika, FireEye.

As can be seen from these examples, investigative reporting on disinformation is funded by a range of actors. Non-profit and non-governmental actors receive a combination of foundation funding, corporate and state funding. Some actors are more institutionally positioned in this respect - for example the Digital Forensic Research Lab at the Atlantic Council is part of a larger entity that receives significant funding from the British, U.S. and UAE governments, and additional backing from Facebook (Lapowsky, 2018). Other entities like Bellingcat are funded by foundations and provide training support and workshops to supplement their core income. Some legacy news organisations are also involved in collaborative investigative work on disinformation which attracts donor funding (e.g. through the International Consortium of Investigative Journalists), while others undertake independent investigations consistent with a mission for journalism designed to hold power to account.

Collaborative and interdisciplinary investigative responses – for example combining the expertise of actors in several of the categories above can heighten the effectiveness of these interventions. For example, Rappler's journalistic investigations in the Philippines have involved partnerships with NGOs, academics and technology experts.

4.2.5 Response case study: COVID-19 disinformation

Due to their more in-depth and resource-intensive nature, and the short timeline of the pandemic, by May 2020 there were fewer published investigative responses to COVID-19 compared to more straight-forward fact-checking and verification efforts. Nevertheless, organisations specialising in investigative responses included outputs from several NGOs[117], news publishers (Evans, 2020), think tanks (EUvsDisinfo, 2020), and joint investigations between academics and independent media (Hollowood & Mostrous, 2020). Topics being investigated include COVID-19 disinformation campaigns launched by state-sponsored media, violent extremist movements, anti-migrant, and far-right networks (ISD, 2020a). These operate across key social communications companies, including Twitter (open posts and direct messaging), Facebook (including profiles, groups, pages, Messenger), YouTube (videos and comments), WhatsApp, and Instagram (open posts and private messaging), despite efforts of these companies to counter the 'disinfodemic'.

Most analysis to date does not involve in-depth investigation by foundations, think tanks or commercial entities, but reporting by news outlets, for example from ProPublica (Kao, 2020) and the *New York Times* (New York Times, 2020) in the U.S., and Rappler (Gaw, 2020) in the Philippines. The overt and continuous spread of disinformation by political leaders during the pandemic has been extensively reported in the media, along with assessments of how statistics are instrumentalised and used to convey misleading impressions.

Another category of investigative responses to COVID-19 disinformation includes guidance on types of disinformation identified to date, such as two policy briefs about the 'disinfodemic' published by UNESCO in partnership with the International Center for Journalists (ICFJ) (Posetti & Bontcheva, 2020). These identified nine types of COVID-19 era disinformation, four main vectors, and ten modalities of response. See also research from

[117] https://rsf.org/en/disinformation

the Reuters Institute on COVID-19 disinformation types, sources and claims (Brennen et al., 2020) which identified political leaders and celebrities as top sources of disinformation.

Within internet communications companies, internal threat mitigation teams, either working independently, or in tandem with other expert actors, were also undertaking investigations into COVID-19 disinformation (Shu & Shieber, 2020). The results disclosed have been piecemeal,[118] and specialist journalists have found them wanting (Turvill, 2020).

4.2.6 How are investigative responses evaluated?

Many actors are transparent on methods and processes and publish publicly accessible reports on their findings. However, explicit evaluations of impact and effectiveness are not publicly available from most of the actors involved in investigative reporting on disinformation. One area of visible results is in the context of industry-driven and collaborative investigations of disinformation campaigns - where specific takedowns of accounts and content related to an investigation occurs on Facebook, Twitter or another social media platform. Similarly, in the context of government-commissioned work, for example into foreign interference in the U.S. 2016 elections, data is directly fed into Congressional hearings.

4.2.7 Challenges and opportunities

Investigative reporting moves beyond individual fact-checking and debunks to produce deeper insights and analysis as well as details on specific campaigns. As this field has matured there is a growing ability to track disinformation actors over time. See, for example, the ongoing tracking of innovations or approaches in reported foreign interference in the U.S. elections 2016 U.S. elections (U.S. Senate Select Committee on Intelligence, 2018) through to campaigns such as IRA CopyPasta (François, et al., 2019).

A challenge to note is that journalists conducting investigations into disinformation are vulnerable to attacks against them, such as online harassment and targeted disinformation about them, as in the case of Maria Ressa at Rappler (Posetti, 2017). A number of internet communication companies have offered a degree of support such as the Facebook - Committee to Protect Journalists safety tips to protect sources and contacts[119], and Google's Project Shield[120]. However, there has been criticism of tardy company responses to complaints of harassment, and to making it the responsibility of the victim to protect themselves by blocking, reporting and deleting rather than the company taking swift action (Posetti, 2020).

As organisations move to codify and quantify the nature of disinformation campaigns, a body of data is developing that enables comparative analysis (as noted above). More organisations also engage in public education alongside intensive report-writing and investigations in order to ensure sharing of good practices and new approaches to countering disinformation. Examples of this include Digital Forensic Research Lab's annual

118 https://en.unesco.org/sites/default/files/unesco_covid_brief_en.pdf
119 https://www.facebookblueprint.com/student/path/188883-journalist-safety
120 https://blog.google/outreach-initiatives/google-news-initiative/our-efforts-help-protect-journalists-online/

Digital Sherlocks methods-sharing conference[121], EU Disinfo Lab's annual conference, as well as Bellingcat's open-source methods and training.

However, in-depth investigations face significant challenges beyond cost and complexity. Most investigations are conducted without access to the complete data sets necessary to fully understand a particular campaign as internet communications companies do not routinely provide this data. Twitter has explained its data disclosure policy in an article by its Head of Site Integrity (Roth, 2019), and Facebook has been criticised by researchers for delays in providing data access but has recently released a larger data set in line with its commitments (King & Persily, 2020). Another issue is the restriction of access to a limited number of researchers, who are also frequently the recipients of large grants from these companies.

Researchers also have limited information and tools to do cross-platform analysis, despite the fact that few organised disinformation (or viral misinformation) efforts are restricted to one single platform. A particular problem is the issue of accessing data on information shared on messaging apps - where disinformation is known to proliferate - which are often end-to-end encrypted for reasons of security and privacy. However, these companies do have access to metadata on traffic and groups, even if they do not have access to specific messages. Access to this information could help investigators to detect patterns of activity by disinformation networks.

In the past four years there has been a heavy initial focus on disinformation deemed to be sponsored by one State in particular. However, as outlined above, recent corporate, academic and investigatory responses are starting to focus on a wider range of States and private/governmental actors involved.

Similarly, as noted above, there are significant gaps in access to information to adequately support civil society, journalism and academia to understand cross-platform as well as messaging-based disinformation campaigns. This represents an opportunity for internet communications companies to collaborate with researchers and civil society organisations with specialist skills in this area on data analysis and policy development.

4.2.8 Recommendations for investigative responses

A number of recommendations for action can be adduced from this chapter for a range of actors. They include:

- All stakeholders could recognise the need to invest in critical, independent investigative journalism as a defensive measure against disinformation, particularly as COVID-19 financial pressures deliver death blows to news outlets around the world and threaten costly investigative journalism initiatives.

- Internet communications companies could provide broader and better access to their datasets to independent researchers studying disinformation, including those who do not receive significant research funding from these companies, in the interests of knowledge sharing to combat disinformation

[121] https://digitalsherlocks.org

Donors and research organisations could:

- Increase investment in interdisciplinary and collaborative investigations, fostering cooperation between academic researchers, commercial data scientists, NGOs and news organisations.
- Fund quick-turnaround disinformation investigations during emergency situations such as the COVID-19 crisis

5 Ecosystem Responses Aimed at Producers and Distributors

5.1 Legislative, pre-legislative, and policy responses

Authors: Trisha Meyer, Clara Hanot, Julie Posetti and Denis Teyssou

This chapter discusses legislative, pre-legislative and policy responses that originate from government actors (legislative, executive, judiciary) and which encompass regulatory intervention to tackle disinformation. These responses cover different types of regulatory action, ranging from inquiries and proposed laws through to legislation and law enforcement. They typically aim at using state power to shape the environment of the production, transmission and receiving of content, affecting either the entire circuit, or specific moments and actors within these.

Disinformation online is tackled from myriad perspectives, including through existing sets of legislation that are not specific to disinformation, but which nonetheless address some aspect of the phenomenon. This chapter cannot cover them comprehensively, but it is worth highlighting some of the means deployed in order to understand the wider legal and policy context in which disinformation-specific government responses develop. Thus the focus here is on legislation and policy strictly related to disinformation, unless it is clear that a legislative/policy measure has been expanded or repurposed to also tackle disinformation.

While institutional and individual self-regulatory approaches are major responses to disinformation, a number of State actors deem it necessary to have regulatory interventions as well. Some of these may be constraining, while others (less often) rewarding. The intention is to provide sufficient disincentives and (less often) incentives to change actors' behaviour. These responses are shaped by national/regional legal traditions, the strength of international legal and normative frameworks, and cultural sensitivities.

In the coercive dimensions of these kinds of interventions, it should be noted that laws applied to disinformation are often vague, which introduces a risk of over-blocking and censoring legitimate expression, including acts of journalism. A further issue is whether existing regulation on harmful expression (for example, on fraudulent claims to sell products) suffices, or whether new regulation is needed and how it can avoid undermining protections for legitimate freedom of expression. Related to this is whether there are effective legal provisions that, in tandem, also ensure that incitement of violent attacks on press freedom and journalism safety (including by disinformation purveyors) is prohibited.

In respect to which some regulatory interventions focus not on restraint, but rather on incentives, an issue is the extent to which there is transparency and equity as a fundamental principle of law. An example is whether there are open and fair systems for regulatory allocation of public funds towards fact-checking, counter-speech (see chapter 5.2 below), or news media, and which ensure that such spending is not abused for political purposes.

Methodology and scope

In order to identify relevant legislative, pre-legislative and policy responses related to disinformation this research has used three resources that cover a range of countries and

approaches as a starting point: the Poynter "Guide to Anti-Misinformation Actions around the World" (Poynter, updated regularly[122]), the Library of Congress report *Initiatives to Counter Fake News in Selected Countries* (Library of Congress, 2019[123]) and the University of Oxford's *Report of Anti-Disinformation Initiatives* (Robinson et al., 2019).

In the analysis of these regulatory responses, the researchers have gone back to the primary sources (laws, policy documents, government press releases, websites, etc) to understand the government initiatives, to the full extent possible. If primary sources proved impossible to find, or where additional information was deemed necessary, secondary sources (news articles, academic reports, legal analyses, etc) were consulted. To be considered reliable, information gained through secondary sources needed to have been found on multiple websites. These secondary sources also led to the identification of additional disinformation-specific government responses.

Some countries propose or have passed legislation unique to disinformation. For others, the proposed amendments or legal basis for tackling disinformation are grounded in other sets of legislation, such as the penal code, civil law, electoral law or cybersecurity law. It is recognised that there are provisions pertaining to disinformation, false information, 'fake news', lies, rumours, etc. in far more sets of legislation than can be covered in one report. Cases have been included where disinformation-related (amendments to) legislation were recently proposed, passed or enforced, or a clear link to disinformation was made in the reporting, discussions and argumentation that led to the proposal, law or its enforcement. Fewer cases have seen countries engage in 'positive measures' as distinct from punitive, and these are discussed in chapter 5.2.

5.1.1 What and who do legislative responses monitor/target?

To understand the tensions and challenges of using legislative and policy responses for freedom of expression, it is worth recalling the right to freedom of opinion and expression, as found in Article 19 of the UN Declaration of Human Rights and echoed in the International Covenant on Civic and Political Rights:

> *"Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."*

Regulatory measures seeking to constrain disinformation should be assessed in terms of the international standards that any restrictions to freedom of expression must be provided by law, be proven necessary to a legitimate purpose, and constitute the least restrictive means to pursue the aim. They should also be time-limited if justified as emergency response measures.

One way of reflecting on how speech is affected by law and policy in online environments, is by assessing responses targeting different *actors' behaviours*. Some responses seek to provide what could be understood as 'positive' measures - necessary conditions for executing the right to freedom of expression. Most measures, however, aim

122 https://www.poynter.org/ifcn/anti-misinformation-actions/
123 https://www.loc.gov/law/help/fake-news/index.php

to deter abusive forms of freedom of expression as defined in law, and thus produce what could be termed 'negative' measures.

Many of these measures are taken with the rationale of protecting citizens. On one side there are steps like data protection rules and media and information literacy policy to give people a level of basic protections and skills to participate in the online environment. At the same time, there are restrictions on expression that cause harm to others, such as incitement to hatred and violence (based on race, ethnicity, gender, or religion), defamation, Nazi propaganda (in specific settings), or harassment and threats of violence. These curbs on speech are justifiable in terms of international standards, although the Rabat Principles of the UN High Commissioner on Human Rights provide important nuance in this regard by setting a high threshold for restrictions.[124] Such constraining elements target inter alia three kinds of behaviours.

Firstly, the range of persons implicated in producing, enabling and distributing content deemed to be harmful are targeted for punishment when *they transgress speech restrictions*. A complication here is whether there can be unintended effects that violate legitimate expression which, even if false (such as in satire) or disturbing (such as in shocking), is not necessarily illegal under international standards. A second and fundamental issue is whether such measures, through design or application, are genuinely to protect the public, or rather to protect particular vested interests such as political incumbents. Additionally, there is the complication that this kind of intended constraint on speech is usually in the form of national-level restrictions that require cooperation from global internet communications companies which have become primary vectors for viral disinformation.

Secondly, competition and consumer protection rules, accompanied by sectoral rules, including phenomena such as laws on misleading advertising, provide the contours of acceptable *economic behaviour* on internet communication companies. However, as chapter 6.3 on de-monetisation responses explains, there is increased questioning within policy circles on whether current rules sufficiently deter economic profiteering from sensationalist and/or false content.

Thirdly, *technical behaviour* is steered through legally formulated cyber-policy seeking to deter use of the internet technologies for malicious intent, such as spam or coordinated information operations for disinformation purposes. Also noteworthy is increased collaboration on topics such as counter terrorism in order to share knowledge and practices among government and technical actors, within legal frameworks on terrorism.

Fourthly, regulatory interventions to channel behaviours of *political actors* include election and political campaign advertising rules.

On the side of enabling, rather than restrictive policy measures, there may be regulatory interventions to increase the availability of information as an alternative to disinformation. These can include enhanced transparency and proactive disclosure practices by *officials*, linked to access to information regimes. They may also include public funds to support *news media*, *fact-checking initiatives*, and *counter-disinformation campaigns by private or public entities*.

[124] Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. https://www.ohchr.org/EN/Issues/FreedomOpinion/Articles19-20/Pages/Index.aspx

Intergovernmental and State-based policy and legal responses to disinformation are cross-cutting and cover all types of actions. Based on the analysis above, four groups can be identified as targets of policy responses.

Firstly, *users considered to be fraudulent and abusive* are at the core of many regulatory responses from governments representing their rationale for action as not only the need to diminish incitements to hatred, violence, and defamation - but also more broadly and problematically, speech that is deemed to be 'false' that is perceived to be prejudicial to national security, international diplomacy, social order, and more (*see #8 in Table 3 below*). On the other hand, some governments also invest in support for *those presenting as ensuring information quality*: fact-checking, counter-disinformation, media and information literacy, and journalism initiatives in order to reliably inform users and empower them to detect disinformation (*see #1,2,3,9,10 Table 3 below*).

Secondly, government initiatives focusing on *internet communication companies* target their economic and technical behaviour. Based on the assumption that online platforms' algorithms enable the viral amplification of disinformation, many regulatory initiatives attempt to place greater obligations on these actors.

In lighter forms of government intervention, internet communications companies are requested to self-regulate and provide public insight into content moderation and political advertising practices and processes. In heavier forms of regulatory action, online platforms and internet intermediaries are required, formally or informally, to de-prioritise, block and take down certain types of content and websites and deregister particular users (*see #5,6,7 in Table 3 belows*).

To some extent, though not often directly targeted, the advertising industry can also be included in this category, as certain policy makers consider the online advertising business model to indirectly enable the financing of disinformation operations (*see #7 in Table 3 below*).

A third stakeholder in the scope of government responses targeting disinformation are *journalists and the news media*. Either by design or unintentionally, many regulatory responses catch journalists and news publishers in the criminalisation of publication and dissemination of false information, despite international protections for press freedom - indicating the need for caveats to shield journalists (*see #8 in Table 3*). In contrast, and as noted above, there are some interventions that have stimulated investment in independent journalism, as well as collaborations between news organisations and communities aimed to strengthen media and information literacy, and third party fact-checking initiatives (*see #1,2,10 in Table 3*), as part of recognising news media's potential role in countering disinformation.

Finally, some government responses target *political actors* (including political parties) themselves, by requiring them to meet new obligations for transparency in online political campaigning, such as the labelling of political advertising (*see #3,7 in Table 3* [125]) and/or by increasing fact-checking endeavours during election periods (*see #1,7 in Table 3*).

[125] See also chapters 4.1 and 7.1

5.1.2 Who do legislative, pre-legislative, and policy responses try to help?

State-based disinformation responses target all involved actors: end users (individuals, communities, audiences, etc), online platforms, advertisers, journalists and news organisations, politicians and political parties, and also domestic and foreign actors perceived to have malicious intent. These regulatory interventions seek to deter what they deem to be abusive forms of expression, with - in the focus of this study - relevance to disinformation, by means of policy and law. Their aim is presented as using 'negative' (i.e. constraining) measures to protect society and its right of access to information by constraining the presence of destructive and harmful disinformation. On the other hand, 'positive' (i.e. enabling) measures aim to affirm the right to freedom of expression by improving the ecosystem through programmes like Media and Information Literacy (MIL) and financial allocations to fact-checkers, media and/or counter-content. However, individual State-based interpretations of rights and responsibilities do not always align with the intent of international legal and normative frameworks designed to support freedom of expression.

'Negative' steps can restrict certain content or behaviour that authorities deem to be fraudulent or otherwise abusive in diverse ways. They focus primarily on moderating the public discourse, under the justification of minimising harm to others, to ensure public health, defence and security but also, at times, for political gain.

Interventions that restrict freedom of expression rights are a notoriously slippery slope, and thus international standards require that they must be provided for by law, be legitimate, proportionate, proven necessary, and the least restrictive means to pursue the stated objective. If they are introduced during emergency settings, they should also be limited by sunset clauses.

On the other hand, "positive" measures targeted at users are aimed, at least in part, at increasing Media and Information Literacy and empowering users via the content they access online. Similarly, they can empower and help enable the news media to investigate, verify, publish and disseminate public interest information.

With respect to the motivating factors, government actions primarily focus on encouraging other actors to tackle disinformation, but they also use the power of legal coercion against actors deemed to be active in the disinformation 'industry'. The theory of change underpinning these kinds of responses will depend on what and/or whom the targets are:

- For *users*, the assumption is that abusive speech can be curtailed through punitive measures, such as fines and arrests. Correlatively, change is expected through increasing the volume of, and access to, credible information, along with awareness-raising among citizens, and Media and Information Literacy programs designed to 'inoculate the herd' against disinformation, so that users are better able to understand and control their own content production/circulation/consumption.

- For *internet communication companies/PR and advertising industry*, the implied theory of change focuses on the role of law and policy in directly - or more often - indirectly reducing the economic and political incentive structures that fuel disinformation. This is also based on the assumption that the companies involved have an interest in thwarting actors who abuse the opportunities that the technology and contemporary business models create. In some cases, the aim is to control the information flows by ensuring that the companies make better use of technology such as AI to deal with issues at scale.

- For *journalists and news publishers*, similar to *users*, the working theory of change is that their publishing 'false' information and speech deemed to be 'abusive' (which, problematically, could capture robust critique as a product of independent journalism) can be curtailed through punitive measures, such as fines, censorship and arrests. The correlative assumption, one aligned with international human rights law, is that change can be effected through support for independent journalism, relying on the belief that the provision of factual and verifiable information shared in the public interest is a precondition for sustainable democracy and sustainable development.

- For *politicians*, the theory of change implicit in related regulatory interventions is that political campaigning, which is largely unregulated online, can be governed by new or updated rules fit for the digital environment. The scrutiny during election periods, through political advertising transparency and increased fact-checking, is considered an incentive for political candidates not to use disinformation as a communication strategy.

The extent to which such perceptions of intervention cause and outcome effect are plausible is discussed in sub-section 5.1.6 below.

5.1.3 What are the outputs of legislative, pre-legislative, and policy responses?

The outputs of state-based responses are reports from inquiries, policy documents (and commissioned research supporting policy development), bills and legislation, and published judgments. In cases where the government takes action, the output would then also include the specific measure taken, such as a fine, an arrest, a campaign aimed to counter what the authority deems as disinformation, or an internet shutdown. In positive measures, there are allocations of resources and capacity-building steps such as for Media and Information Literacy, implementation of access to information regimes, strengthening news media, etc.

5.1.4 Who are the primary actors and who funds them?

Legislative, pre-legislative, and policy work is usually funded by the States, but in some cases - like the internet communications companies - the implementation costs are carried by private entities. Examples are compliance with required transparency of political advertising. This is in line with many commercial enterprises across a range of sectors that have to comply with legislation and policies designed to protect public interests and safety as part of the costs of doing business. At the same time, States may directly finance and execute their own counter-disinformation content campaigns, or media and information literacy programmes.

A multitude of government responses across the globe are covered in this chapter 5.1 as well as in 5.2, 117 responses across 61 countries and inter-governmental organisations. While the objective has been to demonstrate a range of experiences, omissions are inevitable. Most of these policy initiatives are very recent, and many might have been subject to change and review since the time of writing. In addition, inquiries might turn into legislative proposals, legislative proposals might not be adopted, new regulations might arise, amendments might be brought forth, etc. This mapping should therefore be regarded as an evolving tool. The table below also reflects general categories. It does not drill down to more granular identification of issues such as criminalisation of disinformation within the category of legislative responses.

This chapter contains the summary of the research findings. For an entry-by-entry analysis, please refer to Appendix A.[126]

The numbers at the top of the table below resonate with the range of disinformation responses as defined in the study's overall taxonomy, showing links between the legislative/policy responses and the other responses.

1. Monitoring/Fact-checking
2. Investigative
3. National and international counter-disinformation campaigns
4. Electoral-specific
5. Curatorial
6. Technical/algorithmic
7. Economic
8. Ethical and normative
9. Educational
10. Empowerment and credibility labelling

Policy type		Policy response	1. Monitoring/Fact-checking	2. Investigative	3. National and international counter-disinformation campaigns	4. Electoral-specific	5. Curatorial	6. Technical/algorithmic	7. Economic	8. Ethical and normative	9. Educational	10. Empowerment & credibility labelling
Inquiries, task forces and guidelines	**Actor**: *ASEAN, Australia, Belgium, Brazil, Canada, COE, Denmark, Estonia, European Union, India, Indonesia, International Grand Committee, Ireland, Italy, Japan, Mexico, Netherlands, New Zealand, OAS, South Africa, Republic of Korea, Spain, Sweden, Ukraine, UK, U.S.*											
	1	ASEAN's Ministers responsible for Information joint declaration	x		x					x	x	
	2	Australia's Electoral Assurance Taskforce				x	x	x	x		x	
	3	Australia's Parliament Joint Standing Committee on Electoral Matters: Democracy and Disinformation	x			x				x	x	
	4	Belgium's expert group and participatory platform	x								x	x
	5	Brazil's Superior Electoral Court	x		x	x	x					
	6	Canada's parliamentary committee report on 'Democracy under Threat'				x	x	x	x		x	
	7	Canada's Digital Citizen Initiative			x				x	x	x	x
	8	Canada's Critical Election Incident Public Protocol			x	x				x	x	

[126] See Appendix A

Policy type		Policy response	1. Monitoring/Fact-checking	2. Investigative	3. National and international counter-disinformation campaigns	4. Electoral-specific	5. Curatorial	6. Technical/algorithmic	7. Economic	8. Ethical and normative	9. Educational	10. Empowerment & credibility labelling
	9	COE's 'Information Disorder' study	x	x			x	x	x		x	x
	10	Denmark's Elections Action Plan	x	x		x						
	11	Estonia's Cyber Defence League						x				
	12	European Union Code of Practice and Action Plan on Disinformation	x		x	x	x	x	x		x	x
	13	India's social media platforms Code of Ethics				x	x	x			x	
Inquiries, task forces and guidelines	14	Indonesia's war room and 'Stop Hoax' campaigns	x	x	x	x	x			x	x	
	15	International Grand Committee on 'Disinformation and 'Fake News''				x	x	x	x			
	16	Ireland's Interdepartmental Group on 'Security of the Electoral Process and Disinformation'				x	x				x	
	17	Italy's 'Enough-with-the-Hoaxes' campaign and 'Red Button' portal	x	x		x					x	
	18	Japan's Platform Services Study Group				x	x					
	19	Mexico's National Electoral Institute	x			x	x				x	x
	20	Netherlands' 'Stay Critical' campaign and strategy	x	x		x					x	x
	21	New Zealand's parliamentary inquiry into 2016 and 2017 elections				x	x	x	x		x	
	22	OAS' Guide on freedom of expression and disinformation during elections	x	x		x	x	x	x		x	x
	23	South Africa's Political Party Advert Repository and digital disinformation complaints mechanism				x					x	x
	24	Republic of Korea's party task force	x	x								
	25	Spain's government hybrid threats unit	x	x		x						
	26	Sweden's investigation into development of psychological defence authority									x	x
	27	Ukraine's 'Learn to Discern' initiative									x	
	28	UK's House of Commons (Digital, Culture, Media and Sport Committee) inquiry into 'Disinformation and 'Fake News''	x			x	x	x	x		x	

Policy type		Policy response	1. Monitoring/Fact-checking	2. Investigative	3. National and international counter-disinformation campaigns	4. Electoral-specific	5. Curatorial	6. Technical/algorithmic	7. Economic	8. Ethical and normative	9. Educational	10. Empowerment & credibility labelling
	29	UK's House of Commons Foreign Affairs Committee inquiry into Global Media Freedom (sub theme on disinformation)	x	x	x					x		
	30	U.S.' Senate Select Committee on Intelligence inquiry into 'Russian Active Measures Campaigns and Interference in the 2016 US Election'				x						
Legislative proposals	**Actor**: *Argentina, Chile, France, Germany, India, Ireland, Israel, Nigeria, Philippines, Republic of Korea, Sri Lanka, UK, U.S.*											
	31	Argentina's Bill to create a Commission for the Verification of Fake News	x			x	x	x			x	
	32	Chile's proposal to End Mandate of Elected Politicians due to Disinformation				x						
	33	France's Online Hate Speech proposal					x	x				
	34	Germany's Network Enforcement Act update					x			x		
	35	India's proposed amendments to IT Intermediary Guidelines				x	x					
	36	Ireland's proposal to Regulate Transparency of Online Political Advertising				x	x					
	37	Israel's Proposed Electoral Law Amendments and 'Facebook Laws'				x	x					
	38	Nigeria's Protection from Internet Falsehood and Manipulation bill						x		x		
	39	The Philippines' Anti-False Content bill										
	40	Republic of Korea's law proposals				x	x			x		
	41	Sri Lanka's proposed penal code amendments								x		
	42	UK's Online Harms White Paper				x	x					
	43	U.S.' Tennessee State Legislature bill to register CNN and *The Washington Post* as "fake news" agents of the Democratic Party			x							x

Policy type		Policy response	1. Monitoring/Fact-checking	2. Investigative	3. National and international counter-disinformation campaigns	4. Electoral-specific	5. Curatorial	6. Technical/algorithmic	7. Economic	8. Ethical and normative	9. Educational	10. Empowerment & credibility labelling
Adopted legislation		**Actor:** *Argentina, Bangladesh, Belarus, Benin, Brazil, Burkina Faso, Cambodia, Cameroon, Canada, China, Côte d'Ivoire, Egypt, Ethiopia, France, Germany, Indonesia, Kazakhstan, Kenya, Malaysia, Myanmar, New Zealand, Oman, Pakistan, Philippines, Russian Federation, Singapore, Thailand, Vietnam*										
	44	Argentina's Political Party Financing Law				x			x		x	x
	45	Bangladesh's Digital Security Act	x				x	x		x		
	46	Belarus' Media Law						x		x		
	47	Benin's Digital Code								x		
	48	Brazil's Criminal Electoral Disinformation Law				x						
	49	Burkina Faso's Penal Code					x	x		x		
	50	Cambodia's Anti-Fake News Directives					x	x		x		
	51	Cameroon's Penal Code and Cyber Security and Cyber Criminality Law								x		
	52	Canada's Elections Modernisation Act				x	x					
	53	China's Anti-Rumour Laws					x	x		x		
	54	Côte d'Ivoire's Penal Code and Press Law								x		
	55	Egypt's Anti-Fake News Laws					x	x		x		
	56	Ethiopia's False Information Law								x		
	57	France's Fight against Manipulation of Information Law				x	x	x				
	58	Germany's Act to Improve Enforcement of the Law in Social Networks					x	x				
	59	Indonesia's Electronic Information and Transactions Law					x	x		x		
	60	Kazakhstan's Penal Code								x		
	61	Kenya's Computer Misuse and Cybercrimes Act								x		
	62	Malaysia's Anti-Fake News (Repeal) Act								x		
	63	Myanmar's Telecommunications Law and Penal Code								x		
	64	New Zealand's Electoral Amendment Act				x			x		x	
	65	Oman's Penal Code					x	x		x		
	66	Pakistan's Prevention of Electronic Crimes Act					x	x		x		
	67	The Philippines' Penal Code					x	x		x		

Policy type		Policy response	1. Monitoring/Fact-checking	2. Investigative	3. National and international counter-disinformation campaigns	4. Electoral-specific	5. Curatorial	6. Technical/algorithmic	7. Economic	8. Ethical and normative	9. Educational	10. Empowerment & credibility labelling
	68	Russian Federation's Fake News Amendments to Information Law and Code on Administrative Violations					x	x		x		
	69	Singapore's Protection from Online Falsehoods and Manipulation Act					x	x	x			
	70	Thailand's Computer Crime Act					x	x		x		
	71	Vietnam's Cyber Security Law					x	x		x		
Law enforcement and other state intervention		**Actor**: *Bahrain, Bangladesh, Benin, Cambodia, Cameroon, China, Côte d'Ivoire, Egypt, Germany, India, Indonesia, Kazakhstan, Latvia, Malaysia, Myanmar, Russian Federation, Singapore, Sri Lanka, Thailand, Ukraine*										
	72	Bahrain								x		
	73	Bangladesh				x		x				
	74	Benin								x		
	75	Cambodia								x		
	76	Cameroon								x		
	77	PR China					x	x		x		
	78	Côte d'Ivoire								x		
	79	Egypt						x		x		
	80	Germany					x					
	81	India						x				
	82	Indonesia					x	x		x		
	83	Kazakhstan					x	x		x		
	84	Latvia						x				
	85	Malaysia								x		
	86	Myanmar								x		
	87	Russian Federation						x				
	88	Singapore					x					
	89	Sri Lanka						x				
	90	Thailand								x		
	91	Ukraine						x				

Table 3. *Legislative, pre-legislative, and policy responses (mapped against study taxonomy)*

Inquiries, task forces and guidelines

With widespread misinformation and disinformation becoming a growing concern, several countries have set up dedicated task forces and inquiries to monitor and investigate disinformation campaigns. Such task forces have often been launched following disinformation campaigns perceived as a hybrid threat to the country's democratic integrity, or cyber-security. An additional aim of these governmental initiatives is educational, with many including a media and information literacy aspect (*see #9 in Table 3*), such as the Netherlands' 'Stay Critical' strategy (entry 20. in Appendix A) or Canada's Digital Citizen Initiative (entry 7. in Appendix A). In addition, 17 initiatives in this category include fact-checking (*see #1 in Table 3*). It can be highlighted that out of the 30 countries which have set up such inquiries or task forces, 21 have an electoral-specific focus (*see #4 in Table 3*), including a U.S. Senate Select Committee on Intelligence inquiry into interference in the 2016 U.S. Election (entry 30. in Appendix A), Australia's Electoral Assurance Taskforce (entry 2. in Appendix A) and Mexico's National Electoral Institute (entry 19. in Appendix A). Electoral-specific inquiries have the objective to investigate or prevent interference in legislative processes. Because online disinformation is a relatively new phenomenon, most of the initiatives identified are recent and still susceptible to evolution, including as regulatory initiatives.

Legislative proposals

A majority of recent legislative proposals (8 out of 13 analysed) aim to tackle disinformation through curation and the prism of intermediary liability obligations for online platforms regarding misinformation/disinformation or hate speech (*see #5 in Table 3*). This is particularly the scope of France's Fight Against Online Hate Speech Law proposal (entry 37. in Appendix A), Ireland's Proposal to Regulate Transparency of Online Political Advertising (entry 39. in Appendix A) and Israel's Proposed Electoral Law Amendments and 'Facebook Laws' (entry 40. in Appendix A). Similar to inquiries and task forces, the legislative proposals sometimes have an electoral-specific focus (*see #4 in Table 3*), such as Chile's Proposal to End Mandate of Elected Politicians Due to Disinformation (entry 35. in Appendix A). Some other legislative proposals would criminalise the action of spreading disinformation (*see #8 in Table 3*). This can lead to a risk, highlighted on several occasions by human rights activists, of it being used against critical independent journalists.

Adopted legislation

According to this research, by March 2020, at least 28 countries had passed legislation related to disinformation, either updating existing regulations or passing new legislation. The scope of the established legislation varies from media and electoral laws to cybersecurity and penal codes. The regulations either target the perpetrators (particularly individuals and media entities) of what the authorities deem to be disinformation or shift the responsibility to the internet communication companies to moderate or remove specific content, such as the German Network Enforcement Act (entry 61. in Appendix A). In some cases, in particular where disinformation is defined broadly or where provisions are included in general penal codes, there is a major risk of censorship.

Law enforcement and other state intervention

By enforcement of existing or recently adopted laws, a number of State interventions have been justified on the grounds of limiting disinformation. Such actions can consist of fines, arrests or internet and website shutdowns. Enforcement targets individuals, and sometimes journalists and activists; foreign state media considered as disseminating

disinformation (for example Latvia's shutdown of a website linked to another government (entry 88. in Appendix A)); or the internet communication companies deemed as responsible for the massive reach of disinformation (see Facebook fines in Germany (entry 84. in Appendix A)). A number of arrests have been pointed out by Human Rights organisations as arbitrary, and as harnessing disinformation to limit free speech. Internet shutdowns have also been observed to have been used by some governments under a professed rationale of preventing the spread of disinformation, despite such restrictions being blunt (over/under-inclusive) measures that limit access to the full range of information that a society would otherwise enjoy.

5.1.5 How are these responses evaluated?

Many of the impacts of disinformation can be hard to measure comprehensively, and the effectiveness of laws drafted to tackle disinformation are similarly difficult to evaluate. Nonetheless, one example is metrics of action taken by companies against online disinformation (flagging, review, filtering, blocking) which can be considered as criteria for evaluating the application of the law. For example, as a means of evaluating the implementation of the German Network Enforcement Act, platforms report every six months on the action taken on content flagged by users. Partially on this basis, the German government has now proposed updates to the law due assessing the reports having underreported the number of complaints received (Pollock, 2019). A second text revising the initial Network Enforcement Act was expected to be on the table in mid 2020, focusing on the complaint management of the platforms (German BMJV, 2020a; German BMJV, 2020b) (entries 37, 61 and 84. in Appendix A).

It has also become clear that certain laws are difficult to enforce in practice. For example, after the adoption of the French Fight Against Manipulation of Information Law (entry 60. in Appendix A), stakeholders and political candidates sought to demonstrate the limitations of this law. In addition, Twitter initially blocked an official communication campaign from the government to encourage people to vote, arguing it was complying with the law (LeFigaro, 2019). For many small countries worldwide, it is hard in practice to apply laws to international services which do not have significant business or physical presence within the national jurisdiction.

Governments, parliaments and courts can evaluate, and if necessary, revisit and amend existing legislation and policy. For example, the constitutionality of the 2018 Kenya Computer Misuse and Cybercrimes Act has been challenged in court and a judgment was expected in early 2020 (entry 64. in Appendix A). In 2018 Malaysia passed an Anti-Fake News Act. However, after a change of government, the law was repealed on the basis that existing laws (Penal Code, Sedition Act, Printing Presses and Publications Act, Communications and Multimedia Act) already tackle disinformation (entry 65. in Appendix A).

Non-State actors can exert pressure for policy change by publishing their own evaluations and positions on regulatory initiatives. Many civil society groups do in fact provide some evaluations, as do UN organisations such as UN Special Rapporteur on Freedom of Opinion and Expression.[127]

[127] https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/LegislationAndPolicy.aspx

5.1.6 Response case study: COVID-19 disinformation

The COVID-19 pandemic triggered a flurry of state-based actions to prevent and punish acts of potentially life-threatening disinformation (Posetti & Bontcheva, 2020a).[128] Around the world, parliaments, governments and regulators amended or passed laws or regulations enabling the prosecution of people for producing or circulating disinformation, with custodial sentences ranging up to five years (Quinn, 2020). These laws effectively criminalised acts of producing or sharing information deemed to be false, misleading and/or contradicting official government communications about COVID-19. Emergency decrees giving political leaders sweeping new powers were among these measures, along with the application of existing emergency acts to COVID-19 disinformation to enable arrests, fines and jail time for associated offences, such as in South Africa (South African Government, 2020). For example, in January 2020, the Malaysian Communications and Multimedia Commission (2020) detained four individuals suspected of spreading false news on the Coronavirus under Section 233 of the Communications and Multimedia Act.

These measures carried with them the risk of catching legitimate journalism in the net (UK Delegation to the OSCE, 2020). In some countries, producers of independent journalism were arrested and detained, or deported under these laws in the context of States responding to what they deemed to be false information (Simon, 2020; Eljechtimi, 2020). Freedom of expression rights were also affected more broadly due to the challenges of introducing emergency measures in ways that urgently address public health and safety threats, as well as cases of restricting access to official information. Limitations were often not justified, nor in line with the criteria of being legal, necessary, time-limited, and proportionate to the purpose.

Other kinds of policy responses have included support for news media as a bulwark against disinformation. In light of the negative impact of the crisis on the media sector (Tracy, 2020), along with recognition of the corresponding social value of maintaining news outlets, a number of countries took such action.

For example:

- Canada fast-tracked tax relief for media outlets, and put money into advertising specifically to be carried by news outlets (Canadian Heritage, 2020)
- State aid packages or tax exemptions to support news media and media employers were offered in Denmark, Belgium, Hungary and Italy (UNI Global Union, 2020).
- There were mounting calls (Aaron, 2020) for this kind of policy response, qualified by insistence on ensuring transparency, impartiality and independence of any such support mechanisms. Assistance for public service media was also being advocated (Public Media Alliance, 2020).
- A number of NGOs dedicated funds for COVID-19 coverage with state support (UK Government, 2020)

[128] See also the databases of freedom of expression abuses connected to COVID-19 disinformation responses (e.g. 'fake news laws') curated by the International Press Institute (IPI) https://ipi.media/covid19-media-freedom-monitoring/ and Index on Censorship https://www.indexoncensorship.org/disease-control/

5.1.7 Challenges and opportunities

The pace of technological change is a fundamental challenge, as every regulatory action can be quickly outpaced. Broad language can get around this challenge, but at the expense of allowing for interpretations for selective implementation and excessive scope, or for other actors to find loopholes to avoid compliance.

A further challenge is that while there are advantages to dealing with disinformation at the national level, where government initiatives are tailored for a specific political and social context, this does not apply at various supranational levels. This is particularly the case for measures targeting internet communications companies that operate globally. At the same time, it can be difficult for global actors to properly enforce divergent national regulation in the context of networked international information flows.

Some of the measures described in this chapter consist of updating existing legislation to diminish abuses of free expression, and to regulate elections, in order to limit the impact of disinformation on the ability of voters to make informed decisions. Where existing legislation includes protection of freedom of expression and democratic participation, updating or adapting these laws to ensure they can be applied to online disinformation may prevent rushed legislation that does not respect international human rights standards.

Under public and time pressure, legislation is often passed without sufficient debate or transparency, especially in the run-up to elections and in the context of major public health crises like the COVID-19 pandemic. It is noteworthy that some proposed and adopted legislation has been challenged in court, while other bills and acts have been amended or withdrawn in response to such developments.

Moreover, while some governments attempt in good faith to update the regulatory environment to tackle disinformation in the digital age, others have been seen to attempt to control citizens' speech by creating new illegal speech categories, or extending existing laws to penalise legitimate speech. The paradox to highlight here, is that governments that appear to be seeking to control speech for political gain try to legitimise their actions by referring to hate speech regulations and anti-disinformation laws. In other words, disinformation responses risk being used (or justified for use) for censoring legitimate expression - and clearing the field for official disinformation to spread unchecked.

This concern has been increasingly raised by human rights organisations around the world, pointing that such laws have led to abusive arrests of journalists and activists (Human Rights Watch, 2018b; Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b). However, while regulating against disinformation in tandem with safeguarding internationally enshrined rights to freedom of expression can be challenging, there are also opportunities to be noted. For example, when critical independent journalism is empowered as a response to disinformation, governments and private companies can be more effectively held accountable, and policy action can be evaluated and changed as appropriate.

Many legislative and policy responses push responsibility for action onto internet communication companies (especially the big global players), and hold them accountable for the widespread diffusion of disinformation online. But this is sometimes done with insufficient debate and transparency regarding the way measures are then implemented by the companies, and how inevitable risks might be mitigated. Private companies are increasingly required to implement government policy on disinformation, and in essence determine in their implementation the contours of acceptable and unacceptable speech, often with insufficient possibilities of redress for users.

An opportunity is to counter-balance restrictive approaches with enabling measures. Rather than create new expression-based crimes, or to restrict internet access, there are legislative and policy responses which help ensure that information rather than disinformation predominates online. In all cases, there is potential to mainstream assessments of impact on human rights, and on legitimate forms of expression in particular. This covers proposing, passing and implementing state-based responses to digital age manifestations of disinformation.

5.1.8 Recommendations for legislative, pre-legislative, and policy responses

Drawing on the research-based assessment of legislative, pre-legislative and policy responses to disinformation outlined above (and in the accompanying appendix) the following recommendations for action are presented for the consideration of individual States, which could:

- Review and adapt responses to disinformation with a view to conformity with international human rights standards (notably freedom of expression, including access to information, and privacy rights), and make provision for monitoring and evaluation.
- Develop mechanisms for independent oversight and evaluation of the efficacy of relevant legislation, policy and regulation.
- Develop mechanisms for independent oversight and evaluation of internet communication companies' practices in fulfilling legal mandates in tackling disinformation.
- Avoid criminalising disinformation to ensure that legitimate journalism and other public interest information are not caught in the nets of 'fake news' laws.
- Avoid internet shutdowns and social media restrictions as mechanisms to tackle disinformation.
- Ensure that any legislation responding to disinformation crises, like the COVID-19 disinfodemic, is necessary, proportionate, and time-limited.
- Support investment in strengthening independent media, including community and public service media, in the context of the economic impacts of the COVID-19 crisis threatening journalistic sustainability around the world.

5.2 National and international counter-disinformation campaigns

Authors: Trisha Meyer, Clara Hanot and Julie Posetti

This chapter highlights examples of State-based and intergovernmental initiatives aimed at the construction of counter-disinformation narratives, which provide factual information to refute the falsehoods embedded within disinformation narratives. It also discusses whether refutation is an effective disinformation response, based on the latest scientific studies on this topic. Government-run counter-disinformation campaigns have the potential to increase trust and transparency in authorities when they are transparent and serve to enhance dialogue with citizens. An inherent danger, however, is that these mechanisms constitute unidirectional strategic communications initiatives that serve incumbent political interests and also do not address some of the underlying causes of disinformation which would require policies beyond the informational level (such as economic development for marginalised groups or areas). By extension, some counter-disinformation initiatives can risk deepening partisan divides.

5.2.1 What and whom do these responses monitor/target?

Counter-disinformation initiatives launched by national and international authorities target both foreign and domestic disinformation campaigns. While some initiatives do not target a defined type of disinformation, others have a specific focus, such as the European Union External Action Service East Stratcom Task Force which primarily monitors disinformation that it assesses as coming from within countries outside the EU. Some debunking initiatives are actively set up for electoral periods, such as the website led by the Brazil Superior Electoral Court in the run up to the 2018 general elections. Disinformation related to health issues is also a concern that has prompted many dedicated counter-disinformation initiatives, particularly with the COVID-19 crisis.

5.2.2 Who do these responses try to help?

Many of these campaigns and initiatives focus on informing the general public about identified disinformation claims, such as in an electoral context, on a range of policy, natural disasters, and public health and safety concerns amongst others. In addition, the international outreach of such counter-disinformation campaigns can also be designed to preserve or improve the public perception of a country and its government (or regional bloc) on the international scene. These initiatives range from public diplomacy to propaganda. Some of this work provides analysis to military actors, such as the work conducted by NATO StratCom Center of Excellence, which both publishes reports and supports NATO's strategic communications capabilities.

The motivation behind counter-disinformation campaigns is based on refutation. The underlying assumption of those states launching anti-disinformation campaigns, is that debunking and providing accurate factual information to the public will mitigate the belief in, and influence of, non-factual information. There is also the intention to raise public

scepticism based on the provenance of particular messages. A number of these initiatives also go beyond issues of factuality to present narratives and facts in a different light, often thereby hoping to exert geopolitical influence.

5.2.3 What are the outputs of national and international counter-disinformation campaigns?

The work of counter-disinformation initiatives mainly consists of fact-checking activities and dissemination of what is officially considered as authoritative information. The verification is presented online and shared on social media in an attempt to reach the audience on the same platforms where they might encounter disinformation. The debunking can also be directly presented on social media channels, such as the Pakistani @FakeNews_Buster. Such monitoring work might also be presented in reports and extensive analysis to feed strategic communication efforts, such as the work done by the NATO StratCom Center of Excellence and the EEAS East Stratcom Task Force. Additionally, it might be shared with the news media for coverage.

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

The initiatives presented below, collected through research up until May 2020, emanate from governments or international organisations and are thus publicly funded by authorities.

Counter-disinformation campaigns	**Actor:** *ASEAN, Brazil, Cambodia, Canada, China, Democratic Republic of Congo,* **EU/***EEAS, India, Indonesia, Malaysia, Mexico, NATO, Oman, Pakistan, Russian Federation, South Africa, Thailand, Tunisia, UK, UN, UNESCO, WHO*	
	1	Brazil's Superior Electoral Court
	2	Cambodia's TV programme
	3	Canada's programme of activities under the Digital Citizen Initiative
	4	China's Piyao government platform
	5	Democratic Republic of Congo's Ebola mis/disinformation response
	6	European Union EEAS East Stratcom Task Force
	7	India's Army information warfare branch
	8	India's Ministry of Information and Broadcasting FACT check module
	9	Indonesia's CEKHOAKS! debunking portal
	10	Malaysia's Sebenarnya.my debunking portal
	11	Mexico's Verificado Notimex website
	12	NATO StratCom Centre of Excellence
	13	Oman's government communications
	14	Pakistan's FakeNews_Buster' Twitter handle
	15	Russian Federation's Ministry of Foreign Affairs debunking page
	16	Thailand's Anti-Fake News Centre

	17	Tunisia's Check News website
	18	UK Foreign and Commonwealth Office Counter Disinformation and Media Development programme
	19	WHO Coronarvirus Mythbusters campaign*
	20	UN Communications Response (COVID-19)*
	21	UNESCO coronavirus disinformation campaigns*
	22	ASEAN partnership to combat coronavirus disinformation*
	23	EU COVID-19 mythbusting campaign*
	24	South Africa's COVID-19 landing page campaign*
	25	India's WhatsApp coronavirus counter-disinformation* campaign
	26	UK Government's COVID-19 disinformation rapid response unit*

*These initiatives are detailed in the coronavirus case study below

Table 4. *National and international counter-disinformation campaigns*

1. Brazil Superior Electoral Court (2018)

The Brazilian Superior Electoral Court (TSE) launched its own fact-checking and counter-disinformation website (Brazil Superior Electoral Court, 2018)[129] in the run-up to the general elections in October 2018. Reports of disinformation brought to its attention were passed on to the Public Prosecutor's Office and the Federal Police for verification.

2. Cambodia TV Programme (2019)

In early 2019 the Cambodian Ministry of Information launched a weekly live TV programme on the National Television of Kampuchea to counter what it deems to be disinformation (Dara, 2019).

3. Canada Programme of Activities under the Digital Digital Citizens' Initiative (2019)

In 2019, Canada funded a series of initiatives designed to raise awareness about the problem of disinformation and build capacity to combat the problem within broad publics (Canada Government, 2019c).

4. China Piyao Government Platform (2018-)

The Chinese government launched the Piyao ('Refuting Rumours') platform[130], hosted by the Central Cyberspace Affairs Commission in affiliation with the official Xinhua news agency, in August 2018. The platform encourages citizens to report disinformation and uses artificial intelligence to identify rumours. It also distributes state-approved news and counter-disinformation. The platform centralises the efforts of Chinese government agencies to refute what they deem to be disinformation (Qiu & Woo, 2018).

129 http://www.tse.jus.br/hotsites/esclarecimentos-informacoes-falsas-eleicoes-2018/
130 http://www.piyao.org.cn

5. Democratic Republic of Congo Ebola Mis/Disinformation Response (2018-)

In response to the spread of rumours and mis/disinformation about the Ebola virus in the Democratic Republic of Congo, health organisations (WHO, UNICEF, IFRC) collaborated to maintain a database of rumours spread within communities and via social media channels. Because disinformation can complicate the work of medical staff on the ground, the WHO provided fact-checking and risk communication advice for volunteers and frontline personnel in the context of the Ebola epidemic (WHO, 2018). The DRC Ministry of Health also recruited people to report mis/disinformation spread on WhatsApp. These monitoring efforts aim to develop the most appropriate strategy for responding and refuting in person, by radio and via WhatsApp (Spinney, 2019; Fidler, 2019).

6. European Union External Action Service East Stratcom Task Force (2015-)

In March 2015, the European Council tasked the High Representative in cooperation with EU institutions and Member States to submit an action plan on strategic communication. As part of the objective to better forecast, address and respond to disinformation activities by external actors, the task force was set up as part of the European External Action Service to address what it perceived as foreign disinformation campaigns.

For this objective, a small team within the EEAS was recruited to develop what it regarded as positive messages on the European Union in the Eastern Neighbourhood countries. It was also tasked to support the media environment in this region. Finally, the task force analysed the disinformation trend and exposed disinformation narratives, which it saw as emanating mainly from sources outside of the EU. The task force's work on disinformation can be found on their website (euvsdisinfo.eu)[131]. They also operate a Russian language website (eeas.europa.eu/ru/eu-information-russian_ru)[132] to communicate the EU's activities in the Eastern Neighbourhood (EU EEAS, 2018).

The EEAS Stratcom Task Force also operates a 'Rapid Alert System' between the EU Member States, launched in March 2019 as an element of the EU Code of Practice on Disinformation. The mechanism has been first put into use in the context of the Coronavirus crisis (Stolton, 2020).

7. India Army Information Warfare Branch (2019)

The Indian Defence Ministry approved the creation of an Information Warfare branch within the Army to counter what it deems to be disinformation and propaganda in March 2019 (Karanbir Gurung, 2019).

8. India Ministry of Information and Broadcasting FACT Check Module (2019)

Later, in November 2019 the Indian Government announced the creation of a FACT Check Module within the Ministry of Information and Broadcasting. The team will "work on the four principles of find, assess, create and target (FACT)" and will also report disinformation to the relevant government ministries (Mathur, 2019).

131 http://euvsdisinfo.eu/
132 https://eeas.europa.eu/ru/eu-information-russian_ru

9. Indonesia CEKHOAKS! Debunking Portal (2019-)

The Indonesian debunking portal 'CEKHOAKS!'[133] allows citizens to flag disinformation and hoaxes, as well as check which content has been debunked. This website is supported by the Indonesian ministry of Communication and Information Telecommunication, the Indonesian Anti-Slander Society, as well as other government agencies and other civil society organisations.

10. Malaysia Sebenarnya.my Debunking Portal (2017-)

The Malaysian Communications and Multimedia Commission set up a debunking portal 'sebenarnya.my'[134] in March 2017 and accompanying app in March 2018 in order to raise awareness and curb the spread of online disinformation (Buchanan, 2019).

11. Mexico Verificado Notimex Website (2019-)

In June 2019, Notimex, the news agency of the Mexican government, launched its own fact-checking and counter-disinformation website 'Verificado NTX'.[135]

12. NATO StratCom Center of Excellence (2014-)

Based in Riga, Latvia, the NATO Strategic Communication Center of Excellence is a NATO-accredited organisation, formed in 2014 by a memorandum of understanding between Estonia, Germany, Italy, Latvia, Lithuania, Poland, and the United Kingdom. It is independent of the NATO command structure and does not speak for NATO. The Netherlands and Finland joined in 2016, Sweden in 2017, Canada in 2018 and Slovakia in early 2019. France and Denmark were set to join in 2020. The centre analyses disinformation and provides support to NATO's strategic communications capabilities (NATO Stratcom COE, 2019).

13. Oman Government Communications (2018)

The Omani Centre for Government Communications provided training to help the media and communication departments within government institutions to monitor and refute disinformation (Al Busaidi, 2019).

14. Pakistan 'FakeNews_Buster' Twitter Handle (2018-)

The Pakistani Ministry of Information and Broadcasting launched a Twitter handle (@FakeNews_Buster)[136] to raise awareness and refute what it deems as disinformation in October 2018 (Dawn, 2018). A recurring tweet states that "[d]isseminating #FakeNews is not only unethical and illegal but it is also a disservice to the nation. It is the responsibility of everyone to reject such irresponsible behavior. Reject #FakeNews" (@FakeNews_Buster).

133 https://stophoax.id
134 https://sebenarnya.my/
135 http://verificado.notimex.gob.mx
136 https://twitter.com/FakeNews_Buster

15. The Debunking Page of the Ministry of Foreign Affairs of the Russian Federation (2017-)

The Ministry of Foreign Affairs of the Russian Federation has a dedicated webpage to raise awareness of published materials that contain information about the Russian Federation that is deemed to be false.[137] Following the passing of Amendments to the Information Law in 2019, the Russian Federation's media regulator Roskomnadzor was also expected to set up a "fake news database" (Zharov, 2019).

16. Thailand Anti-Fake News Center (2019-)

The Thai Digital Economy and Society Minister set up an intergovernmental 'Anti-Fake News Center' in October 2019 to monitor and refute disinformation, defined as "any viral online content that misleads people or damages the country's image" (Tanakasempipat, 2019b). In coordination with relevant authorities, correction notices are published through the centre's social media accounts, website (antifakenewscenter.com)[138] and the press. The Center has also issued arrest warrants (Bangkok Post, 2019).

17. Tunisia Check News Website (2019-)

A Tunisian fact-checking and debunking website (tunisiachecknews.com)[139] was launched in October 2019. The Tunisian High Independent Authority for Audiovisual Communication (HAICA) supervises the project and works in close collaboration with journalists from public media houses (national television, national radio and Agence Tunis Afrique Presse).

18. UK Foreign and Commonwealth Office Counter Disinformation and Media Development Programme (2016-2021)

In April 2018, in the context of disinformation around the Salisbury poisoning incident (Symonds, 2018), the Foreign and Commonwealth Office (FCO), together with the Ministry of Defence (MoD), Department for Culture, Media and Sport (DCMS) and the Cabinet Office, launched a programme on 'Counter Disinformation and Media Development'. This project is part of a broader set of 'Conflict, Stability and Security Fund programmes in the Eastern Europe, Central Asia and the Western Balkans region'.

The programme provides financial and mentoring support to organisations with the objective to "enhance the quality of public service and independent media (including in the Russian language) so that it is able to support social cohesion, uphold universal values and provide communities in countries across Eastern Europe with access to reliable information." By supporting civil society efforts to expose disinformation, it says that it expects to strengthen society's resilience in Europe.[140]

137 https://www.mid.ru/en/nedostovernie-publikacii
138 https://www.antifakenewscenter.com/
139 https://tunisiachecknews.com/
140 EU Disinfolab responsible for drafting this chapter 5.2 is grantee of the Foreign and Commonwealth Office Counter Disinformation and Media Development programme.

5.2.5 Response Case Study: COVID-19 Disinformation

Counter-disinformation campaigns have been strong elements of both State-based and intergovernmental responses to COVID-19 disinformation. They were rolled out quickly to mobilise online communities to help spread official public health information, as well as debunk content deemed to be false. Partnerships have been forged between various internet communications companies and authorities to provide interactive channels for official content. Measures in this category include campaigns and the creation of special units charged with producing content to counter disinformation.

Examples of these response types deployed to counter COVID-19 disinformation include:

- World Health Organisation mythbusting: In a press conference, a World Health Organisation official declared that "[w]e need a vaccine against disinformation" (WHO, 2020). After the outbreak of the COVID-19 epidemic, WHO set up an official 'Myth Buster's' page[141] to provide reliable information on the disease, as well as an 'EPI-WIN' website.[142] (*This initiative is 19. in the table above*)

- The UN Secretary General launched a UN Communications Response initiative "to flood the internet with facts and science", while countering the growing scourge of misinformation, which he describes as "a poison that is putting even more lives at risk" (UN News, 2020; UN Department of Global Communications, 2020). In May, the initiative was rolled out as "Verified", with the aim being to create a cadre of "digital first responders" to increase the volume and reach of trusted, accurate information surrounding the crisis.[143] (*This initiative is 20. in the table above*)

- UNESCO published two policy briefs deciphering and dissecting the 'disinfodemic' (Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b) which formed part of a broader campaign to counter disinformation and influence policy development at the individual State level. It also produced content in local languages under the rubric of "misinformation shredder".[144] (*This initiative is 21. in the table above*). UNESCO also operated a global campaign called FACTS during the commemorations of World Press Freedom Day on 3 May 2020, audio content was produced in numerous languages for radio stations worldwide, and subsequently launched a further initiative titled Don't Go Viral.[145]

- The UN Global Pulse teams in New York, Kampala and Indonesia are building situational awareness around the outbreak, emergence, and spread of 'infodemics' that can drive efforts across all pillars of the UN, and analytics that identify successful efforts to increase the reach and impact of correct public health information.[146] To this end, they are creating and scaling analytics tools, methodologies, and frameworks to support UN entities to better understand the operational contexts in which they counter the negative effects of COVID-19 in Africa. Based on scientific methodologies, direct support to WHO Africa focuses on providing analytical support and products based on the following methodologies: 1) Short term qualitative and quantitative analysis of digital signals

141 https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters
142 https://www.epi-win.com/advice-and-information/myth-busters
143 https://www.shareverified.com/en; https://news.un.org/en/story/2020/05/1064622
144 https://en.unesco.org/news/faq-covid-19-and-misinformation-shredder-african-local-languages
145 https://en.unesco.org/commemorations/worldpressfreedomday/facts-campaign; https://en.unesco.org/covid19/communicationinformationresponse/dontgoviral; https://en.unesco.org/covid19/communicationinformationresponse/audioresources
146 https://www.unglobalpulse.org/project/understanding-the-covid-19-pandemic-in-real-time/

based on rumours and misinformation provided by field offices; 2) Continuous monitoring based on an adaptive taxonomy which allows identification of rapidly evolving 'infodemics' as well as quantitative evaluation of temporal evolution of particular topics. This includes predictive analytics of rumours and concepts along the lines of size, geographic and channel reach; 3) Sentiment and emotion analysis around particular concepts, including the appearance and escalation of hate speech. This will allow the teams to develop a framework for optimizing the messaging provided by WHO and partners to counter the disinformation.

- The Foreign Ministers of ASEAN and the People's Republic of China met to coordinate their action against COVID-19. In particular, the ministers agreed to strengthen their cooperation in risk communication "to ensure that people are rightly and thoroughly informed on COVID-19 and are not being misled by misinformation and 'fake news' pertaining to COVID-19" (ASEAN, 2020). It has not been precisely described how this cooperation would work in practice. (*This initiative is 22. in the table above*)

- The European Parliament has published guidance on dealing with COVID-19 myths.[147] (*This initiative is 23. in the table above*)

- The South African government has regulated that all internet sites operating within zaDNA top-level domain name must have a landing page with a visible link to www.sacoronavirus.co.za (national COVID-19 site).[148] (*This initiative is 24. in the table above*)

- The Indian Government launched a WhatsApp chatbot designed to counter COVID-19 related disinformation (Chaturvedi 2020). (*This initiative is 25. in the table above*)

- The UK Department for Digital, Culture, Media and Sport (DCMS) set up a dedicated unit to monitor and respond to disinformation on the pandemic, with regular engagement with the internet communications companies (Sabbagh, 2020). This initiative included a 'rapid response unit' which is designed to "stem the spread of falsehoods and rumours which could cost lives" (UK Parliament, Sub Committee on Online Harms and Disinformation, 2020). To complement this effort, the Department of International Development (DFID) supported an initiative to limit the spread of disinformation related to the disease, particularly in South East Asia and Africa. The programme focused on verifying information with help from partner media organisations, such as BBC and sharing reliable news, with help from several selected influencers (UK Department for International Development, 2020). (*This initiative is 26. in the table above*)

5.2.6 How are national and international counter-disinformation responses evaluated?

As many of the initiatives presented in this chapter are quite recent, there is little evidence of meaningful evaluation. At the same time, it appears that the initiatives have also not explicitly embedded monitoring and evaluation activities in their plans which would entail assessment of their intended (and unintended) impact and effectiveness. As they

147 https://www.europarl.europa.eu/news/en/headlines/society/20200326STO75917/disinformation-how-to-recognise-and-tackle-covid-19-myths

148 https://sacoronavirus.co.za

are publicly funded by governments, it is to be presumed that their effectiveness may be assessed internally by governmental services, or externally by civil society and media organisations seeking accountability and transparency. In the context of international operations, the initiatives are evaluated by the Member States that support them. For instance, the April 2018 Foreign Affairs Council (EU Foreign Affairs Council, 2018) "commended the work conducted by East StratCom Task Force" in the context of what it saw as the need to strengthen the resilience of the EU and its neighbours.

5.2.7 Challenges and opportunities

Counter-disinformation campaigns can appear to the target audiences as legitimate and convincing if the institutions initiating them are trusted. Such debunking strategies can also remain within the boundaries of freedom of expression, by refuting content that is not banned as being "false". Where such campaigns are factually grounded and subject to scrutiny, it can be presumed that they are more effective than covert efforts and/or those which are narrative-driven to the point of being propagandistic.

The refutation of 'disinformation' can also be dismissed by critical and disengaged audiences as a public relations exercise for government bodies, rather than a neutral fact-checking exercise. This can, in turn, fuel scepticism and conspiracy theories about State intervention and entrench distrust in State actors, especially those with a history of censorship and propaganda. This is compounded by the risk of governments promulgating their own 'alternative facts' as an exercise in seeding disinformation. Where the same actors themselves might be implicated in the adoption of disinformation tactics, this could be a factor that causes their work of debunking falsehoods to boomerang.

In communications, the 'Barbara Streisand effect' is a widely known theory, according to which the attempt to hide or censor a piece of information can rebound with the opposite unintended consequence of this information going viral in the Digital Age (Masnick, 2003). It is named after the singer Barbara Streisand for her attempt to remove an aerial picture of her property in Malibu which had the opposite effect of drawing more attention to it. This assumption could be tested in relation to the debunking of disinformation as well, including when governmental initiatives are involved.

In cognitive science, this unintended impact is also presented as the "backfire effect", according to which the refutation of information can reinforce the reader's belief in it (Cook et al., 2014; Nyhan & Reifler, 2006). In an analysis of the psychological efficacy of messages countering disinformation, some researchers recommend that when there is a need to repeat a lie to debunk it, it is best to limit the description of it (Sally Chan et al., 2017).

More recent research, however, has not found evidence that retractions that repeat false claims and identify them as false result in increased belief in disinformation. On the contrary, providing a detailed alternative explanation was found to be more effective (Ecker et al., 2019). Some research suggests that debunking should use the modality of a 'truth sandwich' as described by linguist George Lakoff (Hanly, 2018), where the false information is enveloped by true information, and is not given first or last prominence in the narrative.[149] However, further research is needed into differences between governmental debunking and independent debunking.

[149] See detailed discussion of the literature in Chapter 3

One limitation to point out is that refutation only works on identified false claims. Disinformation takes different forms which do not necessarily consist of straightforward false claims, but can involve a decontextualised or misleading application of information to frame an issue, and is often merged with strong emotive resonance.

On the opportunity-side, campaigns led by public authorities can mobilise significant resources - both financial and human - to monitor and fact-check content, and circulate the results. The public character of such initiatives can also lead to public engagement and debate, such as through parliamentary or other oversight mechanisms.

5.2.8 Recommendations for national and international counter-disinformation campaigns

Individual states could:

- Engage more closely with civil society organisations, news organisations, and academic experts to aid development of well-informed campaigns responding to different types of disinformation.
- Consider campaigns designed to raise awareness of the value of critical, independent journalism and journalists in protecting societies from disinformation.
- Invest in research that measures the efficacy of counter-disinformation campaigns.

Researchers could:

- Conduct audience research to test responses to a variety of national and intergovernmental campaign types (e.g. online/offline, interactive, audio-visual) among different groups (e.g. children and young people, older citizens, socio-economically diverse communities, those with diverse political beliefs, those who are identified as susceptible to being influenced by and/or sharing disinformation).

Internet communications companies could:

- Expand financial support for, and heighten the visibility of, intergovernmental anti-disinformation campaigns beyond crises like the COVID-19 pandemic.

5.3 Electoral-specific responses

Authors: Denis Teyssou, Julie Posetti and Kalina Bontcheva

This chapter deals specifically with electoral responses designed to protect voters and the integrity and credibility of elections, through measures that detect, track, and counter disinformation that spreads during election campaigns. Such disinformation threatens democratic processes more generally within a growing number of countries around the world (UNESCO, 2019).

Here, the spotlight is on initiatives launched, either by news media or NGOs, and sometimes by electoral bodies themselves. Their aim is to prevent jeopardising elections and undermining democracy, while preserving universal standards of election integrity applicable to all countries throughout the electoral cycle - from the lead up to elections, during election campaigns, during ballots, and in the aftermath (Norris et al., 2019).

State-based legal and policy responses are detailed in chapter 5.1, while chapter 5.2 tackles counter-disinformation campaigns from States and intergovernmental actors.

Internet Age realities complicate pre-internet normative standards such as those set out in the *Handbook on the Legal, Technical, and Human Rights Aspects of Elections* (United Nations, 1994):

> *"Use of the media for campaign purposes should be responsible in terms of content, such that no party makes statements which are false, slanderous, or racist, or which constitute incitement to violence. Nor should unrealistic or disingenuous promises be made, nor false expectations be fostered by partisan use of the mass media."*

5.3.1 What and whom do electoral disinformation responses target?

The challenges to trust in parts of the news media, combined with the proliferation of user-friendly digital tools that make it easier to create synthetic media that mimics credible journalism, increase the spread of disinformation during election periods (Ireton & Posetti 2018; Norris et al., 2019)[150]. While some falsehoods and myths that spread via orchestrated campaigns are mistaken as factual, the main damage might actually be the systematic erosion of citizens' capacity to even recognise truth. The effect would be to reduce elections to popularity contests which have no need of verified information, eroding the modality of informed voters making rational political choices as a core concept of democratic life.

The kind of disinformation that impersonates legitimate news content is often debunked within a short period of time. However, the purpose behind it is not necessarily to

[150] See also the discussion of trust in chapter 7.1

create a belief based on falsehoods, but rather to "...undermine established beliefs and convictions...to destabilize, to throw suspicion upon powers and counterpowers alike, to make us distrust our sources, to sow confusion." (Eco, 2014). While this observation applies to disinformation across a range of issues (e.g. vaccination, climate change, migration), it can have very direct significance during elections.

For instance, in the context of the 2016 UK EU membership referendum known as the 'Brexit' vote, some researchers argued that voters' exposure to disinformation on social media played a major role in the results (Parkinson 2016; Read, 2016; Dewey 2016). Others, however, pointed out the complexity and polarisation of the political situation as bigger factors (e.g. Benkler et al., 2018; Allcott & Gentzkow, 2017; Guess et al., 2018b), while some highlighted the role of biased coverage in the UK press (Davis, 2019; Freedman, 2016). One Foreign Policy assessment noted the failure of journalistic accountability to professional standards of truth-telling (Barnett, 2016):

> *"Mainstream media failed spectacularly (...) most of UK national press indulged in little more than a catalogue of distortions, half-truths and outright lies: a ferocious propaganda campaign in which facts and sober analysis were sacrificed to the ideologically driven objectives of editors and their proprietors...[Their] rampant Euroscepticism also had an agenda-setting role for broadcasters."*

Another factor here was the failure of objectivity norms within journalism due to the misapprehension that both sides of the debate (i.e. those campaigning for the UK to leave the EU, and those campaigning for it to stay) needed to be given equal weighting, rather than be assessed on the basis of the evidence in the public interest. However, it should be acknowledged that the news media had to navigate a 'pro-leave' political communications strategy designed "to destabilise the discourse while controlling [their] own message based on emotional appeals to voters", which when mixed with disinformation had a powerful impact on democratic deliberations (Beckett, 2016).

Another example highlighting the need to counter election-related disinformation on Facebook and other social media sites was the 2016 U.S. Presidential election. While scholars have emphasised the pre-existing polarisation of American politics, the significance of orchestrated disinformation campaigns (e.g. the Cambridge Analytica scandal) is recognised as a factor in the wider equation (Benkler et al., 2018; Allcott & Gentzkow, 2017; Guess et al., 2018).

Another key concern to be addressed is disinformation associated with political advertising, and its potential to dishonestly influence voters. Such content can be distributed as messages on social networks, within closed chat apps, and in the form of memes, videos, and images to persuade, mobilise, or suppress voters and votes (Wood & Ravel, 2018). Such advertising is designed to affect people's political opinions and voter turnout or suppression. The advertiser pays to produce those effects and can distribute such adverts through microtargeting on social media and search. Some political adverts look like organic content or native advertising, and are also less traceable and thus not easily amenable to counter-narratives. It should also be noted that in many countries the standards applied to political advertising on social media websites are also generally lower than broadcast licensing allows.

Political advertising spending was surging ahead of the 2020 U.S. presidential election, with one digital marketing firm forecasting that the total campaign advertising spend

would jump 63% from the 2016 election, to $6.89 billion (eMarketer, 2020). According to this report, the highly partisan political environment was driving more Americans to donate to their preferred candidates than in previous elections, which in turn was funneling more money into political advertising. While television was predicted to account for the largest share of political advertising (66 percent of the total), digital advertising – with Facebook being the primary platform – was expected to grow more than 200 percent from the previous presidential election, according to the same source. Facebook's ability to offer reach, as well as contentious voter targeting capabilities (Harding-McGill & Daly, 2020), along with its ease of use make it particularly appealing to political advertisers.

These factors combined have given rise to 'sock puppet farms' operated by disinformation agents that span State-linked propaganda units, profiteers, and public relations firms that have begun specialising in creating orchestrated disinformation networks using a host of tactics beyond advertising. These are known as 'black PR firms' (Silverman et al., 2020; Bell & Howard, 2020). There is mounting concern about the role such disinformation purveyors might play in electoral contexts. The danger is that these networks, which also specialise in 'astroturfing', are designed to mimic authentic citizens and organic political movements and therefore generate a veneer of legitimacy which can make their content go more viral than recognisable political advertising.

Another example of deceptive online identities and behaviour emerged in the 2019 UK general election when the name of the Twitter account for the Conservative Party's campaign headquarters (@CCHQPress) was changed to @FactCheckUK, and the accompanying avatar was changed to resemble that of a fact-checking organisation during a televised leaders' debate. Each tweet posted during the debate began with the word "FACT". After the debate, the account name and avatar were changed back. The ultimately victorious Conservative Party defended the act, while Twitter accused the party of misleading the public, a view echoed by the independent fact-checker FullFact (Perraudin, 2019). This weaponisation of fact-checking for political gain during an election campaign underscored the value of such services as tools of trust, while also triggering significant concerns within the fact-checking and journalism communities.

Journalistic actors have responded to these forms of election-related disinformation with investigative reporting and forensic analysis of the data (Ressa 2016; Silverman et al., 2020)[151]. Fact-checking organisations have built on these traditions with electoral specific projects (see below).

State-based responses have involved calls for tighter regulation of political advertising, propaganda networks, and voter targeting in some contexts (Dobber et al., 2019; Kelly, 2020b), but advocated for looser regulation in others (@TeamTrump 2019). The distinction in approaches can be explained in part by the potential for political loss or gain for ruling political parties.

Besides journalists, the other major respondents to electoral disinformation are the internet communications companies themselves. During 2020, a public disagreement erupted between Twitter and Facebook over divergent approaches to fact-checking and identifying disinformation associated with the U.S. President's claims about electoral processes (Smith, 2020b). Fact-checking and flagging his claims as misleading fell within both companies' guidelines on the issue of monitoring and checking electoral disinformation. In May 2020, Twitter took the unprecedented step of attaching a warning

[151] https://www.theguardian.com/news/series/cambridge-analytica-files

label to the relevant tweets (NPR, 2020). This was a move which Facebook CEO Mark Zuckerberg strongly disagreed with, arguing that private companies should not be "arbiters of truth" (Halon, 2020). However, as discussed in chapter 7.1, avoiding being an arbiter of truth does not exclude taking any action against the promotion of clear falsehoods (Kaye, 2020b). In response to Twitter's decision to implement its policies regarding electoral disinformation responses, the U.S. President immediately announced (on Twitter) that he would move to "strongly regulate" or "shutdown" social media companies via an Executive Order (Smith & Shabad, 2020). Civil society organisations focused on freedom of expression condemned the threat, and others said the resulting Executive Order could not be implemented without a change in the law (Article 19, 2020b).

5.3.2 Who do electoral disinformation responses try to help?

Electoral responses are aimed at protecting voters from exposure to disinformation and reducing the likelihood of it influencing their political views and voting intentions in ways that would not have been the case without its impact.

In the context of polls, political advertising (including highly personalised, and individually targeted messaging) has been employed extensively by political parties and candidates with the purpose of influencing voters. For instance, during the 2016 UK EU membership referendum, the #VoteLeave campaign used targeted adverts containing disinformation regarding the weekly cost of Britain's EU membership and Turkey being set to join the EU (Cadwalladr, 2018). In a number of instances, the disinformation was disguised in statistics, which raised complex issues of calculations of costs and benefits. Given that political contests invariably involve selective use of statistics, there are grey areas about when legitimate campaigning blurs into acts of definitive disinformation, although invented statistics are clearly not acts of information. The term disinformation (along with misinformation and 'fake news') themselves can be weaponised to brand particular reality claims as being beyond the pale of accuracy and honesty. These challenges underline both complexity for, and the significance of, responses to electoral disinformation.

Another way in which electoral responses help voters deal with disinformation is to expose the actors behind the problem. For example, many voters do not always know that a major route for targeting them with disinformation is through automated accounts (bots and cyborgs). However, there are well researched cases during the 2016 U.S. presidential elections, the 2016 Philippines presidential election, the UK EU membership referendum, and the 2017 French presidential elections. Political bots, in particular, have been shown as trying to influence voter opinion, e.g. attack political leaders or journalists, although some evidence seems to indicate that bots in certain cases do not change voter intent (Howard et al., 2018b). Nevertheless, it is the case that during elections a large number of (coordinated) bots and sockpuppet accounts were used for spreading disinformation and political rumours (Ressa 2016; Phillips & Ball, 2017; Howard et al., 2018; Gorrell et al., 2018; Howard & Kollanyi, 2016). Exposing such phenomena is part of electoral responses that can sensitise voters to covert disinformation operations and 'dirty tricks' designed to subvert the norms and rules of fair campaigning in a poll.

5.3.3 What output do electoral-specific responses publish?

Outputs of electoral responses to disinformation can include a range of real-time detection, debunks, counter-content, as well as retrospective assessments. They can also entail campaigns linked to voter education, and regulations about electoral conduct.

Electoral responses are treated as a stand-alone response category in this report due to the major impact that disinformation has on democratic processes and citizens' rights during elections. However, this category of responses, due to its very nature, typically involves a combination of monitoring and fact-checking, regulatory, curatorial, technical, educational and other responses, which are separately categorised in the typology presented in this report. They are cross-referenced as applicable in this chapter. Therefore, the outputs from electoral responses essentially equal a subset of the combined outputs produced by these other categories of responses (e.g. election-specific fact-checks, election ad archives).

5.3.4 Who are the primary actors behind electoral-specific responses and who funds them?

a. Political fact-checking in the U.S.

The history of political fact-checking in the U.S. is strongly tied to the coverage of presidential elections, and to the amount of falsehoods spreading during breaking news events. To date, much detailed analysis of the practice of political fact-checking has been focused on the U.S., where the practice is said to have originated (Birks, 2019). In fact, the U.S. fact-checking movement emerged in response to the news media's perceived failure to adequately call out campaign trail falsehoods (Spivak, 2010).

The first independent fact-checking organisation was Spinsanity[152], which was founded in 2001 (Graves, 2013). It was active during the 2004 presidential campaign, producing more than 400 articles. Next, just before the 2004 U.S. presidential election, FactCheck.org was launched as "a nonpartisan, non-profit consumer advocate for voters" which aimed to equally monitor the major political parties, talk shows, TV advertisements, official websites, press releases and media conference transcripts[153]. Another prominent initiative was The Fact Checker, launched by *The Washington Post* prior to the 2008 election (Kessler, 2017). It pioneered a rating system based on one to four Pinnochios.

Another major development in political fact-checking for the 2008 election was the creation of Politifact, the largest independent fact-checking outlet in the United States (Aspray & Cortada, 2019; Drobnic Holan, 2018). They became noteworthy for the quality of their fact-checking and their special Truth-O-Meter rating system (a scale ranging from True, Mostly True, Half True, Mostly False, False, up to Pants on Fire[154]). This Truth-O-Meter became an iconic feature of Politifact (Adair, 2018) which received a Pulitzer Prize in 2009 for its coverage of the 2008 presidential campaign. FactCheck.org was also a nominee (Graves, 2013).

In 2010, Politifact expanded its fact-checking by licensing its brand and methodology to U.S. state-based media partners. Three years later, Politifact launched Punditfact to check the accuracy of claims by pundits, columnists, bloggers, political analysts, the hosts and guests of talk shows, and other members of the media (Hollyfield, 2013).

At present, FactCheck.org, Politifact, the *Washington Post*'s Fact Checker and Snopes are considered to be the most important political fact-checking outfits in the U.S. (Graves 2013; Aspray & Cortada, 2019).

152 http://www.spinsanity.org/about/
153 https://www.factcheck.org/spindetectors/about/
154 https://www.politifact.com/truth-o-meter/

They are facing formidable challenges since, as researchers have argued, there are two media ecosystems in the U.S.: one, "the insular right-wing media ecosystem", which shows "all the characteristics of an echo chamber that radicalizes its inhabitants, destabilizes their ability to tell truth from fiction, and undermines their confidence in institutions"; and another, representing the majority of the news media, that is "closer to the model of the networked public sphere" (Benkler et al., 2018). In this dual media ecosystem, fact-checking websites are perceived as systematically biased by the "insular right-wing" and are generally not trusted or believed by this group (Ibid).

b. Political fact-checking in Europe

Fact-checking as a response to political disinformation started in Europe with a blog launched by UK's Channel 4 News in 2005, to cover a parliamentary election (Graves & Cherubini, 2016). It was followed by similar French press blogs: Désintox from Libération in 2008, and Les Décodeurs from *Le Monde* in 2009. Both were inspired by Politifact and FactCheck.org with the aim of fact-checking politicians and public figures, as well as election campaigns. The British charity FullFact.org began in 2009, with the intent to "fight bad information". That year, it was also joined by the BBC's Reality Check (Birks, 2019). In the Netherlands, the fact-checking project Nieuwscheckers began the same year, within the Journalism and New Media school of Leiden University. Fact-checking has expanded rapidly in Europe, with particular reference to elections. From the 34 permanent outlets active in 20 European countries in 2016, fact-checking at the beginning of 2020 involved some 66 active outlets in 33 countries in the region, according data from Duke's University Reporters's Lab.

Presidential or general elections have often been a catalyst for the extension of the fact-checking movement: either by running a 'real life' experiment, or triggering the establishment of more permanent operations. For instance, French journalists really started fact-checking during the 2012 Presidential election campaign (Bigot, 2019).

In Austria, Spain and Italy, fact-checking went mainstream via TV broadcasting. Austria's public service broadcaster ORF began Faktencheck in 2013 to fact-check politicians on live TV shows in the run-up to the general elections. The same year, in Spain, El Objetivo broadcast a prime time program on fact-checking on La Sexta TV to fact-check politicians amid the Spanish financial crisis. A couple of years later, a similar program made by Pagella Politica was broadcast in Italy on the national channel, TV RAI2.

In 2018, a report from the European Commission's High Level Expert Group on disinformation suggested several strategies in order to overcome disinformation and protect EU elections, as well as elections in Member States, such as enhancing transparency in political advertising, developing tools to empower users, and promoting media literacy. Later that year, the European Commission announced measures for securing free and fair elections to the European Parliament in May 2019 (European Commission, 2018b). Those measures include recommendations for State members to create a national network of relevant authorities to detect and respond to cybersecurity and disinformation threats, greater transparency in online advertising and targeting, and tightening rules on European political party funding.

c. Political fact-checking in the rest of the world

In the Asia-Pacific, Duke University's Reporters' Lab's database registers 47 fact-checking organisations. As elaborated below, regarding elections, disinformation was particularly prevalent in India, Indonesia, the Philippines and in Republic of Korea, while being

substantially lower in comparison in Japan, Singapore, and Australia. In other parts of the world, recent fact-checking initiatives have developed more to debunk disinformation than to verify political claims and discourses, so they are addressed in chapter 4.1.

All political parties in India began using social media in the 2014 election campaigns, with an emphasis on targeting first-time voters (Kaur & Nair, 2018). More recently, WhatsApp has evolved into India's main channel of disinformation (Kajimoto & Stanley, 2019).

A large part of the disinformation debunked in India is political, either pertaining to local disputes or about tensions with neighbouring Pakistan. It is noteworthy that in the legislative assembly election campaign in Delhi in February 2020, members of one political party spread two viral deepfakes videos on WhatsApp with messages targeting a political rival (Christopher, 2020).

In Indonesia, disinformation is often present during important elections, exploiting religious and ethnic fault lines (Kajimoto & Stanley, 2019). The main actors are called 'political buzzers'. They aim to promote their electoral stance, while undermining their rivals' campaigns with hate speech, hyper-partisan discourse, or religious and ethnic-based disinformation.

The Philippines is another country suffering the proliferation of disinformation in online political discourse, especially since the run-up to the 2016 presidential election. The heightened 'indecency' and incivility in political discourse since that period is frequently blamed on so called 'patriotic trolls' and orchestrated online networks (Ressa 2016; Ong & Cabañes, 2018). It has been argued that some fact-checking efforts undertaken by news organisations and NGOs in the Philippines fail to address the underlying causes of disinformation because they do not address the "professionalized and institutionalized work structures and financial incentives that normalize and reward 'paid troll' work" (Ong & Cabañes, 2018). Click farms and the practice of astroturfing, especially on Facebook, have been regularly reported since 2016 in the Philippines.

In Republic of Korea, almost all newspapers and broadcasters launched fact-checking initiatives during the 2017 presidential election (Kajimoto & Stanley, 2019). They aimed to tackle the spread of disinformation, including a collaborative endeavour with academia, SNU Factcheck[155], launched by Seoul National University to enable a fact-checking platform used by 26 news outlets to cross-check disputed information. Other examples in the wider region include the FactCheck Center[156] and MyGoPen, who tackle disinformation on the popular messaging service LINE.

d. Collaborative media responses on elections

Due to the sheer volume of online disinformation and candidate statements in need of fact-checking during elections, a number of media organisations have begun pooling their resources into well-coordinated, collaborative initiatives. Some are national and others are international in nature. The rest of this section discusses some prominent examples.

155 http://factcheck.snu.ac.kr/
156 https://tfc-taiwan.org.tw/

Country-based Collaborative Responses

Electionland was the first **U.S**. joint endeavour in 2016, launched by Propublica with Google News Lab, WNYC, First Draft, Gannett's U.S. Today Network, Univision News, and the CUNY Graduate School of Journalism, to monitor disinformation in social media around the 2016 election day. The project involved 600 journalism students and over 400 reporters located across the U.S. (Bilton, 2016; Wardle, 2017b).

In **Europe**, CrossCheck France (funded by Google News Lab) was among the first collaborative journalism projects on debunking false stories, comments, images and videos about candidates, political parties and all other election-related issues that circulated online during the French presidential election campaign in 2017. It involved more than 100 journalists from 30 French and international media organisations, with some academics and technology companies. In total, 67 debunks were published on CrossCheck's own website, as well as on the websites of the newsroom partners (Smyrnaios et al., 2017). The pioneering collaboration in debunks attracted 336,000 visitors (95% French) (Ibid).

Prior to the UK's 2017 general election, the non-profit First Draft established CrossCheck UK[157], with a dedicated workspace for British journalists providing alerts, facilitating collaborative reporting and investigating suspicious online content and behaviour. In terms of the response categories presented in this report, CrossCheck UK is an investigative response, as the focus is on the disinformation narratives and context rather than on labelling individual claims as true or false. The funding sources for this version of CrossCheck are unclear.

At that time, one of the major challenges faced by such media-focused investigative disinformation responses was the need to establish shared methodology, knowledge and tools[158]. This is where First Draft's contribution was instrumental, alongside the development of innovative tools developed specifically to support collaborative content verification and fact-checking (Mezaris et al., 2019).

First Draft's CrossCheck collaborative methodology was also adopted by the Spanish Comprobado initiative (in collaboration with Maldita.es) to fight disinformation during the country's 2019 general election. In addition to political fact-checking and investigation of disinformation, a new challenge that was addressed was disinformation on private messaging apps (WhatsApp in particular). Comprobado implemented strict quality controls on verification by requiring the approval of at least three of the 16 project members. Based on lessons learned in previous initiatives, Comprobado carefully selected what viral content should be debunked, and how, so as to avoid giving oxygen to disinformation.

In 2018, **collaborative election-focused verification initiatives started spreading worldwide**. One example is the Mexican Verificado 2018[159], led by Animal Politico, Newsweek in Spanish, Pop Up Newsroom and AJ+ Spanish. It aimed to debunk 'fake news' and verify the political discourse during the Mexican 2018 election campaign. It was ground-breaking in scale, as it involved more than 60 media, civil society organisations and universities - all aiming to help citizens decide who to vote for based on confirmed, accurate information. Each report labeled with the Verificado 2018 hashtag was reviewed and supported by the whole network of partners. Verificado 2018 was funded by the

157 https://firstdraftnews.org/project/crosscheck-uk/
158 https://firstdraftnews.org/verification-toolbox/
159 https://verificado.mx/metodologia/

Facebook Journalism Project, the Google Digital News Initiative, and Twitter, as well as the organisation Mexicans Against Corruption and Impunity, and foundations such as Open Society and Oxfam. The initiative won the 2018 U.S. Online Journalism Association Award for Excellence in Collaboration and Partnerships.

The Verificado 2018 initiative was also ground-breaking in terms of the support provided by the internet communications companies. Key to its success were: 1. The companies' provision of access to data about the most shared stories or search engine queries and 2. curatorial measures to promote the verified information. Typically, however, news media and independent fact-checkers lack such comprehensive support and data access from the platforms, which complicates their work significantly.

The Verificado 2018 project is also notable in that it adopted and adapted a collaborative platform originally created by by Verificado19S[160] to manage collaborative citizen response and rescue operations.

In **Latin America** a prominent example is Comprova[161], a partnership of 24 Brazilian media organisations, established for the 2018 elections but still ongoing. The project is coordinated by Abraji (Associação Brasileira de Jornalismo Investigativo) with First Draft. As with many other collaborative fact-checking projects, the Google News Initiative and the Facebook Journalism Project provide financial and technical support to Comprova. Projor, a non-profit organisation focused on issues concerning Brazil's media, was also an early supporter. During the elections, Comprova monitored and verified the veracity of viral information shared by unofficial sources on social media and messaging applications (mainly WhatsApp) (Tardáguila & Benevenuto et al., 2018). WhatsApp monitoring relied on crowdsourced suggestions for content to be verified, leading to 67,000 pieces of information being submitted. This clearly demonstrates the huge volume of potentially problematic political content circulating through these closed messaging apps, and the impossible task that rigorous fact-checking and verifying such content presents.

Another prominent example, this time from Argentina, is Reverso[162]. A massive collaborative project, it was promoted and coordinated by the fact-checker Chequeado, AFP Factual, First Draft and Pop-Up Newsroom, in which more than 100 media and technology companies came together during the 2019 Argentinian presidential election campaign. In order to achieve maximum reach, Reverso debunked (180 articles and 30 videos produced over the six month campaign) were published simultaneously by all partners. The team monitored Facebook, Instagram and Twitter; private messaging apps (mainly WhatsApp); and platforms, such as YouTube and Chequeo Colectivo[163] (a crowdsourcing platform from Chequeado). From an innovation point-of-view and through collaboration with BlackVox[164] the Reverso team also managed to verify fake audio files[165] of candidates shared on WhatsApp[166].

160 https://verificado19s.org/sobre-v19s/
161 https://projetocomprova.com.br/about/
162 https://reversoar.com/
163 https://chequeado.com/colectivo/
164 https://blackvox.com.ar/
165 https://www.clarin.com/politica/reverso-creo-nuevo-metodo-conicet-verificar-audios-virales-whatsapp_0_1638FLWL.html
166 https://www.poynter.org/fact-checking/2019/meet-forensia-a-software-ready-to-debunk-fake-whatsapp-audio-files/

The last prominent Latin American example is Uruguay's Verificado.UY[167] project, in which over 30 partners monitored and debunked disinformation during the Uruguayan presidential elections in October 2019. Training, financial and technological support were provided by First Draft. It focused on two types of verification: rumours spreading on social networks, and statements of politicians and candidates.

In **Australia** and **Asia** respectively, examples include CrossCheck Australia[168] (managed by First Draft), which monitored the 2019 Australian federal election and the collaborative Checkpoint project in India which was operated ahead of national elections there in 2019 (see Chapter 6.1 for details).

In **Africa**, First Draft worked in partnership with the International Centre for Investigative Reporting in Nigeria and 16 newsrooms to establish CrossCheck Nigeria[169] in the run up to the February 2019 Nigerian elections. Support for the project was provided by the Open Society Foundation. It built on knowledge and technology from previous First Draft collaborative initiatives, including Comprova in Brazil and CrossCheck in France. A key feature of this work is the 'CrossCheck' methodology which involves journalists from different newsrooms checking on each other's work to ensure that the principles of transparency, accuracy and impartiality are adhered to.

Another example is the South African Real411[170] ('411' being internet slang for information) project. What is particularly notable is that unlike the previous media- and First Draft-driven examples, Real411 was launched by an NGO (Media Monitoring Africa) and it also involved the South African Electoral Commission. Similar to the other initiatives, it offers an online platform for citizens to report instances of alleged disinformation. This platform, however, incorporates a governmental aspect response, as it is connected to the Directorate of Electoral Offences. Complaints are considered by a panel of experts including media law, and social and digital media representatives. They make recommendations for possible further action for the consideration of the Electoral Commission, including referring the matter for criminal or civil legal action; requesting social media platforms to remove the offensive material; and issuing media statements to alert the public and correct the disinformation. The Real411 site contains a database of all complaints received and their progress. To help distinguish between official and fake adverts, political parties contesting the 8 May, 2019 general elections were asked to upload all official advertising material used by the party to an online political advert repository at www.padre.org.za. This initiative has also been adapted to deal with COVID-19 disinformation.

International Collaborative Responses

EU-wide collaborative responses: FactCheckEU.info[171] was established by the International Fact-Checking Network (IFCN), bringing together 19 European media outlets from 13 countries (the European signatories of IFCN's Code of Principles) to counter disinformation in the European Union ahead of the European Parliament elections in May 2019 (Darmanin, 2019). The core focus was on providing debunks on disinformation or facts about Europe to reduce misperceptions (e.g. islamophobia, immigration). Citizens could submit claims for verification through a web Q&A form which were then picked

167 https://verificado.uy/
168 https://firstdraftnews.org/project/crosscheck-australia/
169 https://crosschecknigeria.org/about/faqs
170 https://www.real411.org/
171 https://factcheckeu.info/en/

up by one of the partners. For maximum coverage, the articles were published in their original languages and translated into English.

This initiative was entirely independent of EU institutions and other governmental actors. The platform was built by the newspaper *Libération* and the web agency Datagif with an innovation grant (U.S.$50,000) from the Poynter Institute. Other costs — primarily the salary of a full-time project coordinator for six months and the costs of translating content — were covered through financial support from Google (€44,000), the Open Society Initiative for Europe (€40,000), and the IFCN (€10,000).

Two other collaborative fact-checking initiatives were launched in parallel: the EU-funded Disinformation Observatory (SOMA, reviewed in chapter 4.1)[172], along with CrossCheck Europe by First Draft[173].

At the time of writing, First Draft's CrossCheck initiative was expanding as a global network of reporters and researchers that collaboratively investigates online content during elections and beyond. Building on the previous campaigns in the U.S., France, Brazil, Nigeria, Spain, Australia and the EU, it seeks to demonstrate that news organisations can work together on a global scale, to produce more effective, efficient and responsible reporting against disinformation.

e. Responses by the internet communications companies

Ahead of the 2020 U.S. presidential election, fears were mounting about exacerbated polarisation, foreign interference, and the rise of new forms of digital content manipulation such as so-called deepfakes: synthetic videos or audio files created through machine learning (see chapter 6.2). Against this background, in 2019 Facebook's vice-president of Global Affairs and Communications, former UK Deputy Prime Minister Nick Clegg, said that "Facebook made mistakes in 2016" but he added that the company had spent the three years since "building its defenses to stop that happening again" (Clegg, 2019). He then enumerated the actions taken by Facebook to crack down on 'inauthentic' accounts – qualified by him as the main source of 'fake news' and malicious content – such as "bringing in independent fact-checkers to verify content" (see chapter 4.1 for an analysis of Facebook's third-party fact-checking network and 7.1 for an assessment of the ethical issues involved) and "recruiting an army of people – now 30,000 – and investing hugely in artificial intelligence systems to take down harmful content".

With respect to false or misleading political advertising, Facebook has been extensively criticised for its policy. The U.S. Sen. Elisabeth Warren accused Facebook of turning its platform "into a disinformation-for-profit machine" and followed up to make a point by publishing a fake advertisement saying: "Breaking news: Mark Zuckerberg and Facebook just endorsed Donald Trump for re-election"[174]. It was reported in October 2019 that Facebook had changed the rules from preventing any advertisements with "false and misleading" content, defined as "deceptive, false, or misleading content, including deceptive claims, offers, or methods," to include a narrower definition prohibiting "ads that include claims debunked by third-party fact checkers or, in certain circumstances, claims debunked by organizations with particular expertise." (Legum, 2019).

172 https://www.disinfobservatory.org/the-observatory/
173 https://firstdraftnews.org/project/crosscheck-europe/
174 https://twitter.com/ewarren/status/1183019880867680256

Facebook further limits its fact-checking of politicians and political parties through guidelines for third party fact-checking partners that state: "posts and ads from politicians are generally not subjected to fact-checking" (Facebook, 2019b.) The guidelines align to "Facebook's fundamental belief in free expression, respect for the democratic process, and the belief that, especially in mature democracies with a free press, political speech is the most scrutinized speech." (See chapters 4.1 and 7.1 for further discussion of these issues). The guidelines indicate: "If a claim is made directly by a politician on their Page, or in an ad or on their website, it is considered direct speech and ineligible for our third party fact checking program — even if the substance of that claim has been debunked elsewhere."

In contrast with Facebook's comparatively hands-off approach to political disinformation, Twitter CEO Jack Dorsey announced that the platform he founded would stop running all political advertisements commencing November 22, 2019. He said that this reflected concerns that "paying to increase the reach of political speech has significant ramifications that today's democratic infrastructure may not be prepared to handle"[175]. As discussed earlier in this chapter, Twitter and Facebook engaged in a public disagreement in mid 2020 over fact-checking and debunking the content published by political leaders, in an incident triggered by Twitter's decision, for the first time, to flag misleading tweets from the U.S. president connected to electoral processes.

In December 2019, Google announced a commitment to "a wide range of efforts to help protect campaigns, surface authoritative election news, and protect elections from foreign interference". (Spencer, 2019). Google said it wanted to "improve voters' confidence in the political adverts they may see on our ad platforms". The company announced changes including limiting election adverts and audience microtargeting to age, gender, and general location (postal code level). They also clarified their advertising policies by explicitly prohibiting "deep fakes", misleading claims about the census process, and adverts or destinations making demonstrably false claims that could significantly undermine participation or trust in an electoral or democratic process.

As part of its election advertising transparency, Google says it provides both in-ad disclosures and an online transparency report[176] (only available for Europe, UK, India and the U.S.) that shows the actual content of the advertisement themselves, who paid for them, how much they spent, how many people saw them, and how they were targeted. "We expect that the number of political ads on which we take action will be very limited—but we will continue to do so for clear violations," the company said. However, Google faced criticism ahead of the U.S. election in 2020 when it refused to remove advertisements from a group accused of voter suppression for falsely claiming that there is a material difference between absentee voting and voting by mail. Facebook, however, agreed to remove similar advertisements from the same group (Stanley-Becker 2020).

f. Regulatory responses to electoral disinformation

Electoral commissions or dedicated government units can also play a key role in fighting electoral disinformation through targeted responses. Examples include actions taken by the Australian Electoral Commission in 2019, including authorisation of electoral communications (AEC, 2019a), and the Spanish 'hybrid threats' government unit which

[175] https://twitter.com/i/events/1189643849385177088
[176] https://transparencyreport.google.com/political-ads/home?hl=en

focuses on cyber security, monitoring, and at times refuting of disinformation (Abellán, 2019).

Naturally, election integrity can be protected through legislative measures. These are discussed specifically in Chapter 5.1 and Annex A, under two dedicated sections - one on legislative proposals and another on adopted legislation.

Electoral commissions and government committees can also provide reliable information on candidates and parties, as well as work with the internet communications companies towards the promotion of such information. For example, the Canadian government created the Critical Election Incident Public Protocol in 2019[177] as a mechanism to notify citizens of election integrity threats as well as inform candidates, organizations or election officials who have been targets of attacks. Another example is the Indonesian Ministry of Communication and Information Technology, which in 2019 organised a 'war room' to detect and disable negative and violating content (Board, 2019). An example of a cooperation response is the approach of the Mexican National Electoral Institute (INE), who signed a cooperation agreement with Facebook, Twitter and Google to limit the spread of electoral disinformation and disseminate practical election information during their 2018 and 2019 elections.

Another important kind of regulatory response targets transparency and integrity of online adverts during election periods. For example, in 2019 the Irish government introduced a legislative proposal to regulate the transparency of online paid political advertising within election periods (Irish Department of the Taoiseach, 2019). A complementary approach is to encourage or to legislate that political parties need to log their online advertising in a public database. In 2019, the South African Electoral Commission for example created the Political Party Advert Repository (padre.org.za) for this purpose.

Responses have also enrolled citizens in helping them discover, report, and act upon electoral disinformation. One example, as already discussed above, is the real411.org portal created in co-operation with the Electoral Commission of South Africa. Another is the Italian government's 'red button' portal, where citizens could report disinformation to a special cyber police unit. The police unit would investigate the content, help citizens report disinformation to the internet communication companies, and in case of defamatory or otherwise illegal content, file a lawsuit (la Cour, 2019).

Another kind of response has been internet shutdowns, although these are widely regarded as disproportionate and even counter-productive to electoral credibility. Some governments have enforced these in the run up to polls saying they are seeking to protect citizens from electoral disinformation and propaganda (Al Jazeera, 2018; Paul, 2018).

There are also some examples of international responses. The European Union adopted an Action Plan on Disinformation, ahead of the 2019 European elections, which aimed to build capacities and cooperation within the EU and among its Member States (European Commission and High Representative, 2018). The European External Action Service also runs a website aiming to provide counter-narratives to disinformation. Another example is the guide to guarantee freedom of expression regarding deliberate disinformation in electoral contexts by the Organization of American States (OAS, 2019). It provides recommendations to a wide variety of actors: legislative branch, judiciary, executive branch, electoral authorities, Internet communication companies, political parties,

[177] https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html

telecommunications companies, media and journalists, fact checkers, companies that trade data for advertising purposes, universities and research centres (OAS, 2019).

5.3.5 How are electoral responses evaluated?

Since many of the electoral-specific responses are actually covered within other response type categories (e.g. fact-checking, curatorial) used to specifically target election-oriented disinformation, the methods and findings from their respective evaluations, as outlined elsewhere in this report, apply fully here.

Regarding transparency in political adverts, Facebook says that "ads about social issues, elections or politics are held to a higher transparency standard on its platform". It adds: "All inactive and active adverts run by politicians on Facebook will be housed in the publicly available, searchable ad library[178] for up to seven years", thereby enabling assessment.

Nevertheless, an Institute for Strategic Dialogue (ISD, 2019) study on the European elections concluded that the Facebook Ad Library "is full of shortcomings". Its classification of adverts is "often haphazard. For example it was accused of having originally wrongly labelled heating engineers in Italy and the Dungeons and Dragons computer game in Germany as 'political' content', while adverts from the far-right German party AfD were missing from the adverts library. In a blog post,[179] Mozilla also complained that Facebook's advertising archive application programming interface was "inadequate", meeting only two of experts' five minimum standards.

The same study (ISD, 2019) argued that internet communications companies are "simultaneously putting freedom of speech at risk, with over-zealous and misguided censorship, while being unable to keep up with many malign campaigns and tactics," the latter also representing threats to freedom of expression. ISD also reported counter-productive measures in Germany, for example, where Twitter's attempts to enable speedy reporting of disinformation appeared to have been gamed by far-right networks, leading to the removal or suspension of the accounts of anti alt-right activists and Jewish-interest newspapers, as well as the victims of harassment, rather than those of the perpetrators.

5.3.6 Challenges and opportunities

Recent research outlines an evolution of disinformation tactics. The already mentioned ISD foresees that "populist parties, far-right cyber militias and religious groups are adapting the tactics more notoriously used by States." The London-based Institute for Strategic Dialogue sees an evolution "away from so-called 'fake news' towards an aggressive 'narrative competition', with the promotion of a 'culture war' dynamic around issues like migration, Muslims in Europe, family vs. progressive values and, increasingly, climate policy" (ISD, 2019). The result is that the connections between political parties and online content are often blurred or fully opaque, the identities of the actors behind messages can be concealed, and there is a lack of transparency around the mechanisms of 'reach'.

In the 2019 EU elections campaign, non-profit activist network Avaaz (Avaaz, 2019) used a crowdsourcing platform[180] to identify new tactics of far-right networks across the

178 https://www.facebook.com/ads/library/
179 https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate/
180 https://fake-watch.eu/

European Union that were adopting the following practices: using fake and duplicate accounts to amplify disinformation spread; abnormal coordination behaviour from specific alternative outlets to share identical content and hate speech; recycling followers with misleading page-name changes; clickbait; and boosting political or divisive agendas through popular entertainment pages. The challenge is to enable such monitoring and exposure during elections, and at scale.

In a 2019 report on the UK General Election (Election Monitoring, 2019), eight organisations highlighted the lack of transparency about the collection and processing of voter data by political parties, and the lack of transparency of political advertising and targeted messaging, including exaggerated and misleading claims. More concerns noted included "opaque funding arrangements" to push "paid content" to voters, bot-like activity in discussions around political parties and policies, spamming of disinformation and conspiracy theories by hyper-partisan actors on Facebook, online harassment of key political figures and journalists, and even the creation of biased polling organisations. The eight signatories, including ISD, Full Fact and the Computational Propaganda Project from Oxford University, called for electoral reform to counter those digital threats to democracy.

The internet communication companies have faced calls to address the challenge of surfacing and promoting reliable information sources during elections, especially as against issues such as deceiving voters (e.g. voter suppression) or undermining trust in the election process (Goldzweig, 2020). As part of their responses to COVID-19 disinformation, the companies have already demonstrated that they have the technical capabilities to do so (Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b) and their challenge is to adapt these to promote reliable information from authoritative sources during elections, such as electoral bodies and/or independent bodies monitoring election integrity.

Another challenge that needs addressing is the funding model for fact-checking and verification organisations, and the sometimes limited transparency associated with these efforts. For example, if an internet communications company controls the fact-checking standards applied to official fact-checks on its sites conducted by third party fact-checkers during elections, and refuses to fund certain content being fact-checked or to apply the results, this may affect the efficacy of fact-checking and how independent and trustworthy such fact-checking efforts are regarded.

Similarly, if fact-checking non-profits and research institutes investigating disinformation content and networks proliferating on social media during elections are directly funded by such companies, what are the implications for their independence, and what safeguards are put in place to ensure funders do not apply undue pressure to these organisations?

These considerations are especially important in light of the great challenge unfolding for internet communications companies to balance their dual responsibilities to uphold freedom of expression rights, while simultaneously consistently flagging, curtailing and blocking disinformation and misinformation during election periods, while facing mounting pressure from powerful political actors to be treated as exceptions to the rules.

Taken together, all these examples highlight the ongoing significant challenges surrounding election disinformation and voter targeting and manipulation. With multiple national elections happening globally on an annual basis, and hundreds of regional and state elections, this presents both major ongoing challenges to the internet communications companies and governments worldwide. But it also brings significant opportunities and impetus to efforts by independent fact-checkers, journalists, media, civil

society, researchers, and national and international organisations to continue - and even expand - their key roles in monitoring, uncovering, countering, curtailing, and evaluating the impact of disinformation.

5.3.7 Recommendations for electoral-specific responses

Given the challenges and opportunities identified above, and the considerable potential harms of disinformation accompanying elections, the following policy recommendations can be made.

Governments and international organisations could:

- Invest in monitoring, measuring and assessing the effectiveness of electoral responses to disinformation.
- Work with internet communications companies to ensure the responses that they initiate are appropriately transparent and measurable, as well as implemented on a truly global scale.
- Encourage internet communications companies to apply the same swift and decisive responses to electoral disinformation as they have to disinformation related to COVID-19.
- Coordinate an initiative to support privacy-preserving, equitable access to key data from internet communications companies, in order to enable independent research on a geographically representative scale into the incidence, spread, and impact of online disinformation on citizens during elections.
- Facilitate and encourage global multistakeholder cooperation and exchange of best practice across continents and States, towards effective implementation of holistic measures for tackling online disinformation during elections.

Internet communications companies could:

- Recognise the significant damage potentially caused by political disinformation, specifically in the run-up to elections (including disinformation in online advertising) and engage in a multi-stakeholder dialogue on the policies and methods they adopt specifically during election periods. These could include temporary restrictions on pre-election political advertising; additional transparency information for political adverts placed during election periods; election-specific policies for promoting reliable information sources; and deployment of additional content moderation and fact-checking resources.
- To deal with cross-platform electoral disinformation, collaborate on the setting of broad industry-wide norms for dealing with electoral disinformation that support democracy and aid self-regulation.
- Collaborate on improving their ability to detect and curtail election disinformation, as cross-platform methods of manipulation are often practiced during elections.
- Apply the lessons learned from responding with urgency to the COVID-19 'disinfodemic' and apply those lessons to the management of political and electoral disinformation.

- Contribute significantly towards funds for fully independent research into manifestations and impact of election disinformation, as well as independent evaluation of the effectiveness of the companies' own disinformation responses, with such initiatives to be managed by arms-length independent funding boards.

- Work together, and under the guidance of the UN Special Rapporteur for the Right to Opinion and Freedom of Expression, along with other independent international experts, to develop a consistent policy approach for dealing with disinformation agents who hold powerful political office while using their sites.

Electoral regulatory bodies and national authorities could:

- Strengthen legislation that helps protect citizens against electoral disinformation (e.g. data protection, freedom of expression, electoral advertising transparency).

- Improve transparency of all election advertising by political parties and candidates through requiring comprehensive and openly available advertising databases and disclosure of spending by political parties and support groups.

- Establish effective cooperation with internet communication companies on monitoring and addressing threats to election integrity.

- Seek to establish and promote multi-stakeholder responses including especially civil society.

- Help educate and empower citizens to detect and report disinformation during elections.

- Improve citizens' knowledge and engagement with electoral processes through civics education and voter literacy initiatives.

- Co-operate with news organisations and specialist researchers in surfacing disinformation and probing disinformation networks.

Media and independent fact-checking organisations could:

- Consider expanding fact-checking during elections to live broadcasts and webcasts, to enable greater reach and impact.

- Carry out research into assessing the efficacy of the different approaches to debunking and containment of disinformation during elections, including responses implemented by regulatory bodies and the internet communication companies.

6 Responses within Production and Distribution

6.1 Curatorial responses

Authors: Trisha Meyer, Clara Hanot and Julie Posetti

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.1.1 Internet communication companies' approaches to content curation

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

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

Facebook and Instagram

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

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

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

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

WhatsApp

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

Google and YouTube

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

Twitter

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

VK

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

Weibo

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

WeChat

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

LINE

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

Snapchat

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

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

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

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

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

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

Table 5. *Curatorial responses from internet communication companies*

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

1. Flagging and review of content

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Promotion and demotion of content

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

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

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

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

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

4. Disabling and suspension of accounts

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

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

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

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

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

5. Transparency in content moderation and sponsored content

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

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

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

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

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

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

6. User involvement

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

7. Appeal

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

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

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

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

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

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

6.1.2 Journalistic curatorial interventions

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

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

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

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

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

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

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

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

6.1.3 What and who do curatorial responses monitor/target?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.1.4 Who do curatorial responses try to help?

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

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

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

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

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

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

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

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

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

6.1.5 What outputs do curatorial responses produce?

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

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

229 https://coralproject.net/about/

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

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

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

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

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

6.1.7 Response Case Study: COVID-19 Disinformation

a. Responses from internet communication companies

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

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

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

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

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

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

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

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

1. *Flagging and review of content*

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

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

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

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

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

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

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

4. *Disabling and suspension of accounts*

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

5. *Transparency in content moderation and sponsored content*

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

6. *Appeal*

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

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

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

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

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

Examples of such journalistic curatorial interventions included:

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

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

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

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

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

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

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

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

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

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

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

6.1.8 How are these responses evaluated?

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

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

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

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

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

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

6.1.9 Challenges and opportunities

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.1.10 Recommendations for curatorial responses

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

Individual States could:

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

International organisations could:

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

Internet communication companies:

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

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

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

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

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

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

The media sector could:

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

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

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

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

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

6.2 Technical / algorithmic responses

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

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

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

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

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

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

6.2.2 Who do technical and algorithmic responses try to help?

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

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

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

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

6.2.3 What output do technical and algorithmic responses publish?

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

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

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

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

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

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

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

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

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

Automated tools for detecting and managing disinformation behaviour

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Third-party detection of disinformation agents, behaviour and networks

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

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

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

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

Third party automated message/content analysis

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

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

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

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

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

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

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

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

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

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

Tools to assist 3rd-party fact-checking

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

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

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

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

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

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

Semi-automated tools to complement content verification

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.2.5 Response Case Study: COVID-19 Disinformation

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

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

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

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

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

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

6.2.6 How are technical and algorithmic responses evaluated?

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

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

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

6.2.7 Challenges and opportunities

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

322 https://www.santaclaraprinciples.org/

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

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

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

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

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

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

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

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

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

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

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

6.2.8 Recommendations for technical and algorithmic responses

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

International organisations and States could:

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

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

Internet communications companies could:

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

Civil society organisations and researchers could:

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

6.3 Demonetisation and advertising-linked responses

Author: Kalina Bontcheva

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.3.5 Response Case Study: COVID-19 Disinformation

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

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

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

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

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

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

6.3.6 How are demonetisation and advertising-linked responses evaluated?

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

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

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

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

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

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

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

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

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

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

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

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

6.3.7 Challenges and opportunities

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

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

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

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

6.3.8 Recommendations for demonetisation and advertising-linked responses

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

Internet communications companies could:

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

Advertising brokerage and exchange networks could:

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

Governments and international organisations could:

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

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

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

7 Responses Aimed at The Target Audiences of Disinformation Campaigns

7.1 Normative and ethical responses

Author: Julie Posetti

This chapter will discuss ethical and normative responses to disinformation executed at international, regional and local levels. These efforts frequently involve public condemnation of acts of disinformation, or recommendations and/or resolutions concerning responses. They extend to initiatives designed to embed values and actions at the individual level that can help counter the spread of disinformation. Because much disinformation may not be illegal (unless it is used for financial fraud, or incitement to violence), there is a wide realm of ethical decision-making by various actors concerning the production, hosting and sharing of fabricated information.

The triangle of norms, ethics and laws can be unpacked in various ways. In this chapter, it is understood that these elements may be aligned, or in tension with each other. Norms and ethics in some cases may run counter to legal frameworks, while personal ethics can involve individuals challenging a particular norm.

7.1.1 What are the aims of ethical and normative responses?

Ethical and normative responses to disinformation primarily operate at the level of shaping norms, thereby reinforcing a shared social assumption that disinformation is a serious threat to society. They are designed to influence individual ethical decisions to identify, counter and deter the production and distribution of disinformation.

This set of responses is not about 'external' protection of the targets or recipients of disinformation, but rather about increasing efforts to prepare people to be active agents in building their own resistance to disinformation. It assumes that the behaviours of those targeted are influenced by norms and ethics, and that the interventions will strengthen these in the interests of helping to 'inoculate' against, and collectively counter, disinformation.

The related expectation is that people are moral, rational and open to 'vaccinating' themselves against viral disinformation. Some evidence suggests, however, that many people choose to believe, endorse and circulate erroneous information that reinforces their beliefs or prejudices, in preference to engaging with accurate, credible content that may challenge them to shift their opinions and add nuance to their identities.

As discussed in chapter 3 (Research Context & Gaps) research in the fields of psychology and sociology has emphasised the complex role and functions of human cognition, belief, and social mores in the disinformation ecosystem. Falsehoods are smuggled into people's consciousness by focusing on beliefs rather than reason, and feelings instead of deduction. The spread of disinformation relies on prejudices, polarisation, partisanship, and identity politics, as well as credulity, cynicism and individuals' search for simple sense-making in the face of great complexity and change (Posetti & Bontcheva, 2020a). This explains why much research indicates, misconceptions can be extremely hard to shift, especially when identifiable facts are shrouded in untruths, even (or, perhaps, especially) when fact-checkers debunk false information. Further, as several sources

have demonstrated, repetition and rhetoric strengthen belief in inaccurate information (e.g. Zacharia, 2019). Other research has concluded that ethical concerns about sharing falsehoods are reduced with repeated exposure (Effron & Raj, 2019). Ethical and normative responses to disinformation should therefore be mindful of these complexities and structured to adapt to them.

The word 'trust' appears dozens of times in this report because many efforts to respond to disinformation are linked to the issue of trust - trust in facts, trust in reputable institutions, and trust in information sources. Trust is implicated as both a vector for disinformation and a potential antidote to it - from the problem of so called 'trust networks' (those networks of family and friends on social media) that propel disinformation (Buchanan & Benson, 2019), to disinformation-laced attempts to undermine trust in critical independent journalism, and encourage cynicism (as distinct from scepticism) and conspiracy thinking about news and truth (Ireton & Posetti, 2018; Vaccari & Chadwick, 2020). Trust is a critical but elusive ingredient in dealing with disinformation. Normative and ethical responses to disinformation impact on the issue of trust by creating a beacon or moral social compass for societal conduct in producing, transmitting, consuming and regulating content.

7.1.2 Who and what are the targets of ethical and normative responses?

The responses in this category are typically aimed at the norms and ethics of targets and recipients of disinformation. Member States of intergovernmental organisations, policy makers, and legal and judicial actors are a primary focus of these interventions. But the broad citizenry, online communities, internet communications companies, news publishers and journalists are also targeted.

These interventions rely on the extent to which those targeted are aligned to international norms on human rights (especially freedom of expression), and are also both able and willing to adhere to codes of ethics, and interested in improving their regulations, policies and practices in response to disinformation challenges.

For example, journalist-oriented initiatives operate on the assumption that journalists have the latitude and the conscience to adhere to codes of ethics (Storm, 2020) and that they are interested in improving the factual accuracy of their coverage in the face of disinformation challenges (Taylor, 2020). They also depend to an extent on which of these standards and norms are embedded within the professional context, and institutionally within news organisations.

Institutional arrangements such as self-regulatory councils are key for underpinning norms and ethics both regarding the media and the internet communications companies. One recent attempt to apply more robust self-regulatory frameworks in this realm is the Facebook Oversight Board (Clegg, 2020; Wong, 2020a). It is a formally appointed semi-autonomous board that will review decisions to remove content (notably, this will not involve informing decisions about what content is kept online in face of complaints). There is no explicit mention of the role of disinformation, misinformation or fact-checking in the Board's charter (Facebook, 2019e), nor what Facebook calls 'coordinated inauthentic behaviour' (i.e. organised misinformation and disinformation), although these may be reasons for content removal. It is not evident what norms and standards will be applied to such determinations if the Board is expected to review such decisions. On the other hand, the newly-appointed Board's Deputy Chair has publicly expressed a desire to "audit" Facdbook fact-checking efforts as part of the Oversight Board's work, stating that

that there are "serious concerns" about political bias in fact-checking and questioning the commitment of fact-checkers to the "facts" (Allen, 2020).

The norms shaping governmental engagement with disinformation are similarly linked to institutional underpinnings such as parliaments, courts and independent communications regulators.

7.1.3 Who are the primary actors and what responses do they produce?

The main actors initiating normative and ethical responses to disinformation are: intergovernmental organisations at the international level (e.g. UNESCO, WHO, UNDP) and regional levels (e.g. EU, CoE, OAS, AU); internet communications companies; news organisations; journalists; and civil society organisations. Below, specific examples of these responses are catalogued and analysed.

a. Intergovernmental responses

At the intergovernmental organisation level, there have been several noteworthy recommendations, statements, and reports produced in an effort to reinforce values and frameworks designed to counter disinformation within the boundaries of international human rights law.

In a significant development in June 2020, a cross-regional statement was issued by more than 130 UN member states and official observers, in the context of COVID-19. This statement said: "It is critical States counter misinformation as a toxic driver of secondary impacts of the pandemic that can heighten the risk of conflict, violence, human rights violations and mass atrocities. For these reasons we call on everybody to immediately cease spreading misinformation… ." The statement further noted "…the key role of free, independent, responsible and pluralistic media to enhance transparency, accountability and trust, which is essential to achieving adequate support for, and compliance by, the general public with collective efforts to curb the spread of the virus". In calling on countries to take steps to counter the spread of such disinformation, the statement advised that efforts should be based on "freedom of expression, freedom of the press and promotion of highest ethics and standards of the press, the protection of journalists and other media workers, as well as promoting information and media literacy, public trust in science, facts, independent media, state and international institutions" (UN Africa Renewal, 2020).

United Nations level responses

Another UN normative intervention is the 2017 'Joint Declaration On Freedom Of Expression and Fake News", Disinformation and Propaganda' (OSCE, 2017). This declaration was issued by 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.

This joint statement, produced in collaboration with the civil society organisations Article 19 and the Centre for Law and Democracy, came in response to a rash of legislation from multiple states seeking to address the disinformation crisis by prohibiting the

publication and dissemination of certain content. It seeks to address both the causes and consequences of disinformation (including both disinformation-fuelled attacks on the news media by state actors, and the rush to regulate against disinformation) through the framework of international human rights law, emphasising enshrined freedom of expression rights. The statement indicates that the signatories are:

> *"Alarmed at instances in which public authorities denigrate, intimidate and threaten the media, including by stating that the media is "the opposition" or is "lying" and has a hidden political agenda, which increases the risk of threats and violence against journalists, undermines public trust and confidence in journalism as a public watchdog, and may mislead the public by blurring the lines between disinformation and media products containing independently verifiable facts."*

Recognising the potential for so called 'fake news legislation' to infringe on freedom of expression rights, in particular through inadvertently (or by design) curtailment and suppression of legitimate journalism, it also emphasises that:

> *"...the human right to impart information and ideas is not limited to "correct" statements, that the right also protects information and ideas that may shock, offend and disturb, and that prohibitions on disinformation may violate international human rights standards, while, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or State actors."*

The objective of such statements is to sensitise UN Member States about their responsibilities under international human rights law, and to encourage adherence as a way of dissuading both the use of disinformation as a tool to intimidate or regulate the news media and other publishers of public interest information as a means of limiting freedom of expression. The target audiences of such interventions also include policy makers, the news media, and by extension, the broader public.

Associated approaches to reinforcing normative values and ethical standards adopted by UN agencies include UNESCO's #MILCLICKS campaign and its 'Journalism, 'Fake News' and Disinformation' handbook (Ireton & Posetti, 2018). The former initiative seeks to foster Media and Information Literacy[356] (MIL), through Critical-thinking, Creativity, Literacy, Intercultural, Citizenship, Knowledge and Sustainability (CLICKS).[357] It is aimed at young audiences, and is designed to foster critical engagement with information online - a cornerstone of medium and longer term responses to disinformation - promoting the notion of #ThinkBeforeSharing. The normative practice being encouraged is accountability for communications, as well as informed and ethical reflection about how individuals engage with content. The UNESCO handbook, meanwhile, is aimed at embedding ethical, accountable and critical approaches to combatting disinformation within journalism education and training. The handbook adopts an ethical framework for journalism's defence against disinformation: "Ethical journalism that values transparent practice and accountability is a vital piece of the armoury in the battle to defend facts and truth in an era of 'information disorder'." (Ireton & Posetti, 2018). It further elaborates:

356 https://en.unesco.org/themes/media-and-information-literacy
357 https://en.unesco.org/MILCLICKS

> *Professional standards for ethical and accountable journalism are an important defence against disinformation and misinformation. Norms and values providing guidance to people doing journalism have evolved over the years to give journalism its distinctive mission and modus operandi. In turn, these uphold verifiable information and informed comment shared in the public interest. It is these factors that underpin the credibility of journalism. As such, they are woven into the fabric of this handbook. (Ireton & Posetti, 2018)*

Regional level responses

Policy initiatives, charters of obligations, inquiries and targeted research from the European Commission and the Council of Europe have contributed to a comprehensive attempt to reinforce normative and ethical responses to disinformation in Europe.

The European Commission promotes the normative understanding that disinformation can "cause public harm, be a threat to democratic political and policy-making processes, and may even put the protection of EU citizens' health, security and their environment at risk." (European Commission, 2019). It outlines its policy approach and intent regarding efforts to combat disinformation in its online policy repository, with objectives summarised thus:

> *The exposure of citizens to large scale disinformation, including misleading or outright false information, is a major challenge for Europe. The Commission is working to implement a clear, comprehensive and broad set of actions to tackle the spread and impact of online disinformation in Europe and ensure the protection of European values and democratic systems. (European Commission, 2019).*

This approach has been informed by collaborative scholarship and expert consultations, including the work of the EU's High Level Expert Group on 'Fake News' and Online Disinformation. In its final report (Buning et al., 2018), the Group made a series of recommendations that emphasise the values of privacy, professional ethics, and social responsibility.

One initiative to flow from the European Commission's normative policy approach (European Commission, 2018a) is an Action Plan Against Disinformation (European Commission, 2018e) which is designed to deal with legal acts of disinformation and is couched in terms of geopolitical threats and the need to reinforce European democratic values: "This Action Plan was a response in 2019 to the European Council's call for measures to 'protect the Union's democratic systems and combat disinformation, including in the context of the upcoming European elections'."

Another action-oriented outcome focused on ethics is the European Commission's Code of Practice on Disinformation (European Commission, 2018c), which was published in late 2018 with the assertion that: "This is the first time worldwide that industry agrees, on a voluntary basis, to self-regulatory standards to fight disinformation." Signatories now include Facebook, Google, Twitter, Mozilla and Microsoft, along with eight advertising trade associations (European Commission, 2018d). Stated objectives of the Code include transparency in political advertising, although there is no reference to accuracy or fact-checking associated with political advertising. This is relevant to ongoing debates

connected to political advertising during elections, which have seen calls for the introduction of 'truth in political advertising' standards.[358]

The Council of Europe commissioned a foundational research report which emphasises the role of professional ethics and norms in combatting what it defines as 'information disorder' (Wardle & Derakhshan, 2017). The report offers a range of recommendations for states, technology companies, the news media, civil society and funders. Beyond disruption to democratic elections, the report identified the biggest concern demanding attention as: "...the long-term implications of disinformation campaigns designed specifically to sow mistrust and confusion and to sharpen existing sociocultural divisions using nationalistic, ethnic, racial and religious tensions." (p. 4) This points to the need for responses to disinformation that recognise the risks at the intersection with hate speech and seek to reinforce norms and values like racial and gender equality, and religious tolerance.

The 2017 "Joint Declaration On Freedom Of Expression and 'Fake News', Disinformation and Propaganda", referenced above, was also signed by regional intergovernmental organisations representing Latin America and Africa, along with OSCE. The OSCE Representative on Freedom of the Media has also reiterated: "...at all times, and especially in difficult times, blocking or banning media outlets is not an answer to the phenomenon of disinformation and propaganda, as it leads to arbitrary and politically motivated actions. Limits on media freedom for the sake of political expediency lead to censorship and, when begun, censorship never stops. Instead, the answer lies in more debate and media pluralism".[359] Additionally, the OSCE has supported country-specific workshops designed to embed freedom of expression norms in responses to disinformation while practically equipping Member States to respond to disinformation (OSCE, 2017).

b. Civil Society responses

Many civil society responses to disinformation involve initiatives that seek to reinforce democratic values and human rights frameworks that support norms like freedom of expression, access to information, privacy and gender and racial equality. Several of these interventions, operating at the intersection of disinformation and hate speech, are detailed in section d. below.

Many of the examples of Media and Information Literacy initiatives from civil society organisations identified in the next chapter are also designed with strong normative and ethical components at the core. Such initiatives seek to stimulate grassroots ethical responses to disinformation.

One notable civil society initiative designed to address information pollution is Reporters Without Borders' (RSF) Forum on Information and Democracy[360] based upon an international declaration endorsed by 38 countries. This initiative evaluates norms and architectures of global communications networks, investigates companies' actions, makes recommendations, facilitates regulation and self-regulation, commissions research and supports journalism.

358 See discussion below

359 https://www.osce.org/fom/319286

360 https://informationdemocracy.org/; disclosure: the author of this chapter is a member of the steering committee of the Forum's Working Group on Infodemics https://informationdemocracy.org/working-groups/concrete-solutions-against-the-infodemic/

c. Responses from the internet communications and news industries

From the internet communications companies to news organisations, a range of normative and ethical responses to disinformation can be catalogued.

Twitter decided to ban political candidate advertising from its site ahead of the 2019 UK elections, with company CEO and founder Jack Dorsey announcing via a tweet: "We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought." (Dorsey, 2019). Google followed suit a couple of weeks later, replicating Twitter's commitment to prevent micro-targeting of users for politically-themed adverts. Additionally, Google promised to ban 'deepfakes' and what it termed "demonstrably false claims" to try to protect the integrity of elections and support trust in democratic processes (Wong, 2019a).

As a result, Facebook came under mounting ethical pressure to address its policies pertaining to misinformation and disinformation connected to political advertising and speech on its site (see chapter 4.1) - brought into sharp focus by the Cambridge Analytica scandal[361] - after it decided not to apply fact-checking standards to certain types of political advertising (Eisenstat, 2019; Stewart, 2019). Facebook considered restricting the micro-targeting of users by political actors (Glazer, 2019). However, the company ultimately announced that it would not curtail such micro-targeting, and that no action would be taken to prevent politicians from making false claims in their posts, nor in paid advertising, ahead of the 2020 U.S. election (Romm et al., 2020). Under this policy (see chapter 4.1), the company further excluded certain types of political advertising content from the fact-checking work which it contracts out (meaning therefore that it also does not label this kind of content as false and misleading) (Hern 2019a; Van Den Berg & Snelderwaard, 2019) . However, the company did proceed with new protocols in the U.S. that meant it could ask its fact-checking partners to assess the truthfulness of non-political advertising on Facebook (Hern, 2019b).

Facebook's normative argument is that, in general, it is inappropriate for a private company to be an arbiter of truth in the case of political advertising (Gilbert, 2019). In a 2019 blog post, Facebook's Vice-President for Global Affairs and Communications, Nick Clegg, argued that freedom of expression is "an absolute founding principle for Facebook" (Clegg, 2019). As noted by UN Rapporteur on Freedom of Opinion and Expression, David Kaye, avoiding being an arbiter of truth should not exclude Facebook from taking any action against clear falsehoods (Kaye, 2020b). The normative debate in practice is balancing the company's interpretation of freedom of expression with actual limitations on expression set out in the company's community standards, and how these limits compare to those permissible for states to make under international human rights law. The result is controversy over whether cases violate Facebook's own community standards or raise issues of restriction under international standards (which the private sector is expected to respect, according to the UN's principles agreed in the Ruggie Report[362]). An example is conspiracy theories, which in principle are tolerated on the service, unless these are deemed to contain false or misleading content that can cause imminent harm. There was, however, evidence of a more restrictive approach emerging in mid 2020, when Facebook removed nearly 800 pages and groups, and restricted approximately 2000 Instagram accounts in connection with the QAnon conspiracy theory. (Facebook, 2020b)

361 https://www.theguardian.com/news/series/cambridge-analytica-files
362 https://www.business-humanrights.org

Ethical concerns about Facebook's approach to fact-checking political advertising motivated hundreds of the company's employees to argue in a letter to management that: "Free speech and paid speech are not the same thing." They claimed that policies on avoiding fact checking advertisements from politicians, political parties and their affiliates "are a threat to what FB stands for". They stated that the policy does not protect voices, but instead "allows politicians to weaponize our platform by targeting people who believe that content posted by political figures is trustworthy." (New York Times, 2019)

It is important to note, however, that Facebook policy still allows the rejection of direct speech or advertising by political candidates, incumbents, political parties and their affiliates if it amounts to an immediate threat to safety "in the real world", or if it contravenes the company's voter suppression policies (Facebook, 2019d). For example, on March 30th, 2020, Facebook and Instagram removed videos of Brazilian president Jair Bolsonaro for spreading disinformation on the coronavirus and therefore violating the platforms' terms of use. Those terms do not allow "misinformation" that could cause physical harm to individuals, said Facebook (BBC News, 2020b). However, these standards are not applied uniformly internationally. For example, posts quoting U.S. President Donald Trump speculating on bleach as a potential treatment for COVID-19 were not removed (Suárez, 2020).

Although Facebook CEO Mark Zuckerberg was cited stating that promoting bleach as a cure for coronavirus was the kind of "misinformation" that would be removed immediately - because of "imminent risk of danger" - the company said that Trump's statement did not violate the policy because he did not specifically direct people to ingest bleach. Since then, Facebook has removed a video in which the U.S. President claimed children were "virtually immune" to coronavirus (Kang and Frenkel 2020). The issue of Facebook applying its standards differently around the world has been recognised by former senior Facebook policy manager Richard Allan, who explained differences in treatment in terms of "risk" related to the proximity of a country to the U.S. and its size. (Suárez, 2020). In September 2020, Buzzfeed published extracts from a memo by a former Facebook data scientist who claimed that outside of Western countries, the company regularly abrogated its responsibility to deal with political disinformation with the potential to cause physical harm. She cited instances in multiple developing countries. (Silverman, Mac and Dixit, 2020).

Related concerns were also raised in a UK House of Lords report which assessed that "Facebook have purposefully hobbled their third-party fact checking initiative by exempting all elected politicians and candidates for office from being fact checked." (House of Lords, 2020).

Deciding when content is opinion or fact when these are closely intertwined in a given item requires, inter alia, an ethical judgement call. As discussed in chapter 4.1, this highlights policy loopholes whereby disinformation may not be labelled as such, or fact-checking labels denoting falsity are removed by the company, because falsehoods are bundled with opinion (which Facebook policy regards as largely exempt from fact-checking), thereby creating conundrums for what constitutes an appropriate response at an ethical level. For example, Facebook has removed fact-checking labels applied by third party fact-checkers to content deemed to be opinion (Penney, 2020; Pasternack, 2020; Grossman & Schickler, 2019). There are also reports of pressure being applied by the company to third party fact-checkers in reference to the fact-checkers' assessment of opinion and 'advocacy' content, and fact-checkers being wrongly accused of bias with regard to labelling scientific disinformation, with very limited transparency (Pasternak, 2020).

While Facebook has long-running formal fact-checking partnerships with many reputable news organisations and NGOs,[363] several of which have described a mission-driven motivation for participating (Funke & Mantzarlis, 2018a), the initiative has attracted ethical critiques from some journalists. Those actively engaged in third party fact-checking who feel that the collaboration clashed with professional norms have been among these critics (Levin, 2018). A number of fact-checking partners have ultimately pulled out of the arrangement in the midst of debates on professional ethics connected to the operation of Facebook's Third Party Fact-Checking Program (Lee, 2019a). Among them was U.S.-based anti-hoax website Snopes. One of Snopes.com founders indicated that the ethical challenges were among the reasons for withdrawing (Green, 2019). Facebook's fact-checking partner in the Netherlands, Nu.nl, also withdrew from the project[364]. The non-profit outlet took a values-based decision to quit the collaboration in disagreement with Facebook's adoption of an ethical position to exempt political advertisements (with some exceptions) from its fact-checking (Hern, 2019a; Van Den Berg & Snelderwaard, 2019).

Another example of a news organisation demonstrating competing norms is the BBC's complaint about a Facebook advertisement which used a decontextualised and misleading clip of Political Editor Laura Kuenssberg appearing to endorse the Conservative Party's Brexit strategy. This presented significant reputational and ethical challenges for a public broadcaster that holds up political neutrality as one of its core values. Facebook banned the advertisement several days after receiving the complaint, with the justification that this was a copyright breach (Mays, 2019). At the time it was banned, GBP 5000 had been spent on the advertising campaign which had appeared in news feeds about 250,000 times (Who Targets Me, 2019).

More recently, Facebook moved to thwart politically affiliated publishers masquerading as local news sites from claiming exemption from the company's political advertisement authorisation process (Fisher, 2020). This followed publication of research from the Tow Center for Digital Journalism that revealed over 1200 cases of political groups posing as local news sites to publish propaganda in the U.S.. (Bengani, 2020).

Additional ethical and normative responses to disinformation have come from some news organisations putting disinformation combat at the core of their editorial strategies. For example, a 2019 study (Posetti et al., 2019a) identified a 'mission-driven' approach to combating disinformation from three Global South news organisations: Rappler in The Philippines, the Daily Maverick in South Africa, and The Quint in India. Each of these news organisations identified a commitment to reinforcing democratic principles, defending media freedom, and adhering to the core ethical tenet of 'speaking truth to power' in response to state-sponsored disinformation networks and foreign influence agents that they believed were destabilising their democracies. Additionally, they sought to model these norms for their audiences as a means of motivating the ethical responsibility to eschew disinformation practices, including attacks against journalists laced with 'lies'. One example of this approach is audience-focused campaigns from Rappler, encouraging the community to join them in opposing online hate connected to orchestrated disinformation campaigns which targeted Rappler and its CEO-Executive Editor, Maria Ressa (Posetti et al., 2019b). These were operationalised online using hashtags like #NoPlaceForHate, #IStandWithRappler and #HoldTheLine, the objective being to demonstrate a shared ethical commitment to combating disinformation within online communities and opposing state-based disinformation campaigns as being antithetical to cultural and social norms and mores.

[363] See detailed analysis in Chapter 4

[364] Facebook's Third Party fact-checking programme has since relaunched in the country with two partners: AFP and DPA

In 2020, in calling on Facebook to assume moral responsibility to act in response to disinformation, Ressa cited the UN's conclusion that Facebook had played a "determining role" [365] in what the UN has described as a "a textbook example of ethnic cleansing"[366] against the Rohingya in Myanmar through its facilitation of both disinformation and hate speech (Posetti, 2020). Facebook later acknowledged that "we weren't doing enough to help prevent our platform from being used to foment division and incite offline violence," and said it had updated its policies to "now remove misinformation that has the potential to contribute to imminent violence or physical harm."[367]

The South African National Editors' Forum (SANEF) has also played a normative role in highlighting the dangers of disinformation and working closely with Media Monitoring Africa to confirm a commitment to the eradication of disinformation[368]. Initiatives connected to this collaboration include the disinformation reporting portal called Real411.

Another collaborative response to disinformation from the journalism community came during the 2019 World Media Summit (an initiative of China's Xinhua news agency, which now involves 13 international media partners, including Reuters, BBC, and AP). The Summit reportedly reached a consensus on disinformation: "To ensure the authority and credibility of media are upheld, media have the mission to fight against disinformation; false information should be clarified without delay; and fake news should be boycotted by all... . The reporting and spreading of fake news violate journalistic ethics and damage the interests of the general public" (Xinhua, 2019).

d. Anti-hate speech initiatives

Where disinformation intersects with hate speech - such as racism, misogyny and bigotry - normative and ethical responses are often triggered. These span initiatives from civil society organisations, and intergovernmental agencies through to interventions from celebrities. One such celebrity is comedian-actor Sacha Baron Cohen, whose speech on social media-fuelled disinformation and propaganda to an Anti-Defamation League conference on antisemitism and hate in November 2019 sought to get the companies involved to take action against disinformation endangering religious and ethnic minorities (Baron Cohen, 2019).

There are also interventions from research institutes and NGOs seeking to provide normative guidance through development of frameworks designed to embed values-based approaches to managing hate speech as it manifests as a feature of orchestrated disinformation campaigns. One example of such an intervention is the International Foundation for Electoral Systems' exploration of the links between hate speech and disinformation, and provision of a normative framework for programming interventions (Reppell & Shein, 2019).

Another example is an RSF report 'Attack of the Trolls' that covers the online abuse of journalists - particularly at the intersection of misogyny and disinformation. It sought to raise awareness and activate responses designed to reinforce press freedom norms online (RSF, 2018) Similarly, the FOJO Media Institute's #JournoDefender initiative[369] focused on combatting online misogyny as it intersects with disinformation fuelled-attacks designed

365 https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf
366 https://news.un.org/en/story/2017/09/564622-un-human-rights-chief-points-textbook-example-ethnic-cleansing-myanmar
367 https://about.fb.com/news/2018/11/myanmar-hria/
368 https://sanef.org.za/disinformation/
369 https://journodefender.org/

to undermine democracy. This initiative was underpinned by research conducted in multiple countries (FOJO: Media Institute, 2018).

Journalists and news organisations have themselves sought to reinforce values of gender equality by investigating disinformation campaigns involving misogynistic elements. For example, Rappler Editor and CEO Maria Ressa cites a commitment to the principles of 'speaking truth to power' and 'shining a light' as reasons she chose to speak out publicly about her experience of being brutally harassed online in retaliation for investigative journalism that exposed reportedly government-linked disinformation networks in the Philippines. (GIJN Staff, 2019)

Finally, UN Special Rapporteurs have signalled online hate-speech deploying disinformation tactics against female journalists. Five UN Special Rapporteurs issued a joint statement in 2018 calling on the Indian Government to protect Indian journalist Rana Ayuub who was bombarded with death threats as part of a misogynistic disinformation campaign which used 'deepfake' videos and fake accounts to misrepresent her and expose her to risk (UN Human Rights, 2018).

7.1.4 Response Case Study: COVID-19 Disinformation

Ethical and normative responses include public condemnation of acts of disinformation, or recommendations and resolutions aimed at thwarting these acts because of the life-threatening character of the pandemic. Such responses include statements from UN special rapporteurs, WHO officials, and national leaders. Additionally, there have been examples of calls for reinforcing ethical conduct within journalism, and for internet communications companies to do more. These responses have often taken the form of published statements, speeches or articles designed to move others to stop sharing disinformation, to reinforce freedom of expression norms during the crisis, and to adapt ethical standards to address new challenges in responses to what two UNESCO-commissioned policy briefs responding to the COVID-19 disinformation crisis framed as the 'disinfodemic' (Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b).

Examples include:

- A World Press Freedom Day statement from UN Secretary General Antonio Gutteres reinforcing the normative role of professional journalism in the information ecosystem as a bulwark against disinformation. This statement also asserted the ethical and legal obligations of UN Member States regarding press freedom rights (and journalism safety mechanisms) in the context of responses to COVID-19. (UN Secretary General, 2020)

- A joint statement from International experts including David Kaye, UN Special Rapporteur on the right to Freedom of Opinion and Expression; Harlem Désir, OSCE Representative on Freedom of the Media, and Edison Lanza, IACHR Special Rapporteur for Freedom of Expression: "Governments must promote and protect access to and free flow of information during pandemic". (UN Human Rights, 2020a)

- A report to the UN Human Rights Council from UN Special Rapporteur on the right to Freedom of Opinion and Expression David Kaye which explicitly appealed to the moral and ethical obligations and responsibilities of Member States in reference to their COVID-19 responses (Kaye, 2020a). The report states that it is a "plea to all

Governments to treat those within their jurisdictions ... with the dignity and respect demanded by international human rights law."

- Calls from senior editors, journalists and media academics to stop live broadcasting politicians who disseminate disinformation during speeches and press conferences, due to the difficulty of fact-checking and debunking in real-time. (Thomas, 2020)

- Unprecedented decisions by internet communications companies to edit or remove recordings of political leaders deemed to be spreading disinformation about COVID-19. (BBC, 2020b)

- As noted above, the crisis triggered more than 130 United Nations member countries and official observers to urge that all steps to counter COVID-19 should be based, inter alia, on respect or freedom of expression and press freedom.[370]

- The African Commission on Human and Peoples' Rights issued a press statement on a "human rights based effective response" to the pandemic. This reiterated the obligation of States to ensure that the measures adopted comply with the principle of legality, are necessary and proportional to the objective of safeguarding public health. Such measures include those intended "to dispel misinformation and myths about COVID19 and to penalize the dissemination of false information on risks of COVID19".[371]

7.1.5 How are ethical and normative responses evaluated?

There is very limited evidence of any kind of evaluation associated with ethical and normative responses to disinformation, in part because of the methodological difficulty of such an exercise. One contributing factor is that embedding ethics and norms within societies, or stimulating commitments to international human rights principles is a highly collaborative process and it is close to impossible to determine which actor, or which particular message, was more or less transformative.

The relevant UN Special Rapporteur monitors Member States' handling of disinformation in reference to their adherence to international human rights norms like freedom of expression, and issues assessments to the UN Human Rights Council on that basis (UN Human Rights, 2020b). But there is no known evaluative process that seeks to directly attribute the development of norms and ethics within societies to such interventions. Some case references however do show impact in certain contexts.[372]

For example, a statement from the UN Secretary General (UN Secretary General, 2020) highlighting the fundamental importance of ensuring that counter-disinformation measures introduced by Member States do not undermine the principles like press freedom, is typically amplified by the news media (Apelblat, 2020) and reinforced by civil society organisations' efforts to embed norms like 'access to information' (Article 19, 2020). However, beyond media measurement exercises by commercial service providers,

370 https://www.un.org/africarenewal/news/coronavirus/cross-regional-statement-%E2%80%9Cinfodemic%E2%80%9D-context-covid-19
371 https://www.achpr.org/pressrelease/detail?id=483
372 https://www.lefigaro.fr/flash-actu/la-bolivie-abroge-des-decrets-anti-desinformation-controverses-20200515

there is not systematic and at scale publicly-available research about the extent of dissemination and amplification of these kinds of statements.

7.1.6 Challenges and opportunities

These normative and ethical interventions can be comparatively simple and affordable to implement and they can work as counter-narratives that appeal to individuals' moral compasses, or reinforce alignment with values like anti-racism or anti-misogyny. A problem, however, is when moral compasses and societal norms are not linked to the principles of access to information, freedom of expression, press freedom, and privacy - as enshrined in international human rights law. There are many attempts to normalise the expression and dissemination of false and misleading content that is potentially harmful.

One of the most significant risks associated with state-based responses to disinformation is posed by legal and regulatory approaches that go against the international norms of freedom of expression (including its corollary press freedom) and privacy.

As highlighted by the rush of responses to the disinfodemic that accompanied the COVID-19 pandemic, a disinformation crisis can lead to changes in what is accepted as normal, such as the suspension or weakening of human rights protections (Posetti & Bontcheva 2020a; Posetti & Bontcheva 2020b). Further, even though restrictions on the right to seek, receive and impart content can be legitimate under international standards, when these are for reasons of public health, it remains the case that they need to be in law, as well as necessary and proportional to the purpose.

In such circumstances, responses like 'fake news' new laws that effectively criminalise journalism can exceed these standards, and also go on to become entrenched as new norms. It is therefore a challenge to ensure that all interventions are anchored within the legal and normative frameworks of human rights, and particularly freedom of expression (including press freedom and access to information) and privacy.

Ultimately, legitimate normative and ethical responses to disinformation can be de-legitimised by the individuals, organisations and States who disagree with the intention behind them, in the same way that credible journalism can be misrepresented as 'fake news' by those seeking to avoid being held to account.

Many actors highlighting these issues seek to address the challenge of a risk of downgrading human rights standards by empowering the public (and their elected representatives) to recognise that such interventions against disinformation (where they are justified during emergencies like the COVID-19 pandemic) should inter alia be of time-limited duration. However, the impact of these messages depends on persuading those with power to tack closely to these standards.

The main opportunity in the category of ethical and normative responses to disinformation is to reaffirm and remind people about norms around access to information and freedom of expression. In the COVID-19 crisis, it can be underlined that these norms are not only about fundamental rights, but also significant tools for mitigating impact and tackling disinformation.

Immediate normative steps to counter disinformation can also be taken with an eye to promoting long-term normative and institutional impacts in terms of international standards. For example, news reporting on disinformation responses can explain the

importance of assessing these initiatives against international human rights laws, and the normative and ethical frameworks that support them. Such explanatory journalism could aid accountability on the part of governments and corporate actors, as well as help embed understanding of the role of these values and standards in areas beyond dealing with disinformation.

7.1.7 Recommendations for normative and ethical responses

The challenges and opportunities identified above, and their significant implications for freedom of expression, provide a basis for options for action in this category of responses.

Recommendations for action in this chapter include strengthening the institutional underpinnings for freedom of expression norms, as components of disinformation responses. In this regard:

International organizations could:

- Conduct follow-up evaluation of the circulation of, and engagement with, normative statements as well as assessment of the actual impact of ethical codes, such as operated by internet communications companies and news media that are relevant to disinformation issues.

Individual states could:

- Ensure institutionalised multi-stakeholder governance of internet communications companies, covering transparency and the range of policies on disinformation in the context of content curation.
- Embed human rights impacts assessments within responses to disinformation from executive or legislative branches of government, especially those which risk overreach (e.g. the expansion of 'fake news' laws in the context of the COVID-19 pandemic).

Internet communications companies could:

- Commit to values that defend vulnerable communities and groups, including from threats in multiple languages, and ensure that all countries in which they operate are served by measures adopted to combat disinformation.
- Engage diverse stakeholders in developing policies that support ethical decision-making concerning disinformation content - including if it should be removed.
- Increase capacity to deal with disinformation at scale, especially in countries in conflict, and provide swift responses to actors targeted by this disinformation, as well as redress opportunities in regard to decisions on how relevant content is treated.

- Strengthen their normative role regarding freedom of expression by ensuring regular independent review of their disinformation-related policies and implementation, and the human rights impacts thereof.

- Recognise that an ethical commitment to freedom of expression does not preclude a range of decisive actions on political disinformation that is likely to cause significant harm - such as where it threatens lives, public health, the institutions of democracy, or electoral integrity.

- Enhance transparency and disclosure of data about practical processes around managing disinformation.

Media actors could:

- Ensure that they adhere to the highest ethical and professional standards to avoid becoming captured or associated with disinformation purveyors.

- Invest in investigative journalism focused on exposing disinformation networks and explaining the risks of disinformation to their audiences and the importance of resisting it in the public interest, as a means of building trust while also pursuing truth.

- Increase the capacity of independent press councils to monitor and address disinformation (including when it spreads through news media channels) and disinformation responses (especially as they affect freedom of expression) as part of their ethics oversight role.

Researchers could:

- Use audience research methods to measure the influence and impact of messaging aimed at developing ethics and values that help inoculate against disinformation, or undertake qualitative research into normative evolution and behavioural change focused on disinformation defences.

- Study Media and Information Literacy initiatives to assess the impact on participants' behaviours and sense of personal accountability regarding the need to counter disinformation.

7.2 Educational responses

Authors: Kalina Bontcheva, Julie Posetti and Denis Teyssou

Educational responses are aimed at improving citizens' media and information literacy and promoting critical thinking and verification in the context of online information consumption, as well as journalism training and tools designed to strengthen fact-checking, verification, and debunking.

Of particular relevance are critical thinking, news and advertising literacy, human rights awareness, identity issues, understanding of algorithms and data, and knowledge of the political economy of communications (including economics, psychology and sociology around the production, targeting and spread of disinformation).

This section provides an overview of different kinds of educational responses, distinguished by cataloguing the organisations that design and deliver them and identifying the targets of these responses. In conclusion, they are assessed as to how they address disinformation in relation to educating learners about the fundamental value of freedom of expression, and explain the difference between mobilising and interpreting different facts on the one hand (which would not constitute disinformation), and on the other, when false or misleading information is mobilised and interpreted (which is the essence of disinformation).

7.2.1 What and who do educational responses target?

Media and information literacy and critical thinking initiatives are widely regarded as key 21st century skills, required by citizens to more effectively discern and counter online disinformation. As noted in a report by the Broadband Commission: "Traditional school curricula tend to prioritize the accumulation of knowledge over the application of knowledge, and many school systems fail to adequately train students in digital citizenship and literacy." (Broadband Commission, 2013).

The notion of Media and Information Literacy (MIL) as UNESCO uses it, includes a range of competencies concerning the consumption, production and circulation of content. Under this umbrella are knowledge and skills covering fields such as critical thinking, content production; news literacy; advertising literacy; film literacy; political economy of communications; algorithmic literacy; privacy literacy; and intercultural communication (Berger, 2019). UNESCO also operates with a concept of Global Citizenship Education[373] (GCED), which includes competencies around identity and values. Together, these represent a "playbook" that can help empower participants in digital communications to deal with disinformation in a range of ways. Educational initiatives in the wide field of MIL may be formal and informal, and spread across a range of social institutions from schools through to cities, transportation systems, as well as media and social media.

UNESCO's wide range of target competencies emphasises the comprehensive breadth required for MIL efforts to be successful. While many efforts tend to focus on news and

[373] https://en.unesco.org/themes/gced

verification literacies alone, the strongly interconnected topics of algorithmic, advertising, and privacy literacy are very rarely addressed. The notion of "digital literacy " is variably elaborated as to what competencies it aims to prescribe. As argued by some civil society organisations[374], it also is very important to educate children (and adults) about how personal data are collected and shared online for commercial gain; the hidden dangers of online profiling and targeting; algorithms and their biases; and user privacy online. Thus an important gap in a number of MIL toolkits and programmes is in the lack of coverage on the concept of data literacy. Meanwhile, data literacy in the face of disinformation links closely to the issue of digital citizenship (Carmi et al., 2020).

Addressing these MIL challenges through long-lasting, effective educational responses is a key part of the puzzle, since research has found in some instances that the main amplifiers behind viral disinformation are human users (Vosoughi et al., 2018). The key questions, then, are why do citizens 'fall for' online disinformation, what motivates them to share it (even if they are aware it is untrue), and what is the impact of online disinformation on their offline behaviour (e.g. does it affect their voting in elections)? Particularly in the context of COVID-19, many citizens are being duped and are propagating online disinformation, leaving them unable to understand and implement scientifically-grounded preventive measures. People are dying as a result of complacency (Karimi & Gambrell, 2020), or resorting to false 'cures' (Embury-Dennis, 2020).

Both scientists (e.g. Corbu et al., 2020) and fact-checkers (Vicol, 2020) have been studying the question of what makes citizens believe and spread false or misleading content . Age, education, and repetitive exposure to disinformation have all been confirmed as important factors (Vicol, 2020), with adults over 50 and those without higher education being particularly at risk. Another important factor is *confirmation bias*, i.e. people's tendency to read and believe content which conforms to their existing worldviews (Nickerson, 1998; Corbu et al.,2020; Nygren & Guath, 2019). According to a study by Gallup and the Knight Foundation (Knight Foundation, 2018), people generally share information that they trust and do so primarily for social or personal reasons. Moreover, an individual's online news and information sharing and commenting behaviour is influenced by the behaviour of their typically like-minded online social connections - referred to as *homophily* in scientific studies (Tarbush & Teytelboym, 2012). Receiving content from trusted sources such as friends and families adds credence to the credibility of this content.

Possibly linked to all this, researchers have found that "social networks and search engines are associated with an increase in the mean ideological distance between individuals" (Flaxman et al., 2018), i.e. lead to polarisation. These findings hold across many countries (Kelly & François, 2018).

Experimental research has also shown that when polarised online communities are exposed to disinformation which conforms to their preferred narratives, it is believed and shared (Quattrociocchi et al., 2016). Consequently, when such users and communities are exposed to debunks or opposing opinions, these may either have little effect or can even strengthen their pre-existing beliefs and misconceptions. Moreover, a recent FullFact survey showed that homophily motivated 25% of UK adults to share content even though they believed it to be made up or exaggerated (Vicol, 2020).

Researchers from the YouCheck! MIL project have also found evidence of overconfidence (Nygren & Guath, 2019; Nygren & Guath, 2020) and a `third person effect' (Durach,

[374] https://5rightsfoundation.com/our-work/data-literacy/

2020; Corbu et al., 2020), where people rate their own capabilities to detect online disinformation too favourably compared to the abilities of others.

Taken together, this evidence demonstrates the key importance of developing effective MIL and GECD responses to disinformation.

7.2.2 Who do educational responses try to help?

MIL and GCED are widely regarded as key skills that enable citizens to discern online disinformation more effectively. Where citizen-oriented surveys have been carried out, however, evidence has emerged consistently that the majority of citizens are lacking these essential skills. For instance, a 2018 Eurobarometer survey (Eurobarometer, 2018) in the 28 EU member states established that only 15% of the respondents felt very confident in identifying online disinformation. Other surveys focused specifically on the citizen's ability to distinguish factual from opinion statements, where a Pew Research Center study (Mitchell et al., 2018) has shown that (on average) only 26% of Americans were able to recognise factual news statements, with the number rising to 33% for younger Americans. In addition, a RISJ report (Newman, 2019b) findings have indicated a global tendency to conflate poor journalism with disinformation and 'fake news'.

This has motivated the emergence of initiatives aimed at improving media, digital and data literacy, and critical thinking across all ages (from school children, through to retirees). Data literacy in the face of disinformation links closely to the issue of digital citizenship (Carmi et al, 2020).

Complementing these efforts are initiatives and resources, aimed at educating journalism students and professional journalists in the most up-to-date tools, methodologies, and resources for verifying, investigating and debunking online disinformation. These are often developed and facilitated by leading journalists, journalism educators, researchers, and civil society organisations. Frequently, these efforts are also highly collaborative.

7.2.3 What output do educational responses publish?

Outputs Aimed at Improving Citizen's MIL and GCED

One class of media literacy initiatives relies on learning through **online games**, i.e. teaching citizens media literacy and critical thinking through participation in a game. This is an engaging way for people (not just school children) to gain knowledge and experience. One example is the Drog initiative[375] which has brought together academics, journalists, and media experts to build an online game - GetBadNews. The game aims to educate people about the various tactics employed in online propaganda and disinformation campaigns. Another educational game is Fakey[376] by the University of Indiana, which asks players to share or like credible articles and report for fact-checking suspicious ones. The BBC has developed the iReporter[377] interactive game (Scott, 2018), which gives young players the role of a journalist who needs to report on news without falling prey to disinformation. Another notable example is the multilingual YouCheck!

[375] https://aboutbadnews.com/
[376] https://fakey.iuni.iu.edu/
[377] https://www.bbc.co.uk/news/resources/idt-8760dd58-84f9-4c98-ade2-590562670096

Detectives[378] fake news game, which is available in English, French, Spanish, Romanian, and Swedish. For older teens (15-18 years old), the global International Factchecking Network has produced a role-playing card game (currently in English, Italian, Portuguese, and Spanish),[379] with students playing newsroom journalists covering a controversial referendum, marred by online propaganda and disinformation. Other examples have been collected by the American Press Institute (Funke & Benkelman, 2019). The connection between MIL and games has also been recognised by UNESCO and selected governments in a pioneering Games Bar session held in late 2019.[380]

There are also more traditional, **school-based approaches** to media literacy, which are targeting pre-teens and teens, just as they start taking interest in social media, news, and politics. A prominent government-led response comes from Finland, where the public education system has media literacy classes as standard (Henley, 2020). This is reported to have made Finnish citizens well prepared to recognise online falsehoods. Elsewhere, media organisations and civil society groups are filling the gap in state-led provision in schools. Examples include the school media club run by the NGO African Centre for Media and Information Literacy (AFRICMIL)[381] and the MENA student-oriented MIL activities of the Media and Digital Literacy Academy of Beirut (MDLAB).[382] Another example is Lie Detectors[383] - a non-profit initiative in Belgium and Germany, which puts journalists in the classrooms to interact directly with pupils and teach about news literacy and news verification practices. There is a similar initiative in France led by journalists from Le Monde (Roucaute, 2017) and the Redes Cordiais journalist-led initiative in Brazil.[384]

A more global initiative comes from the BBC, who with the help of the British Council, are providing global media literacy resources for schools around the world (BBC, 2020a). Elsewhere, politicians have spearheaded a programme to teach media literacy to high school children, an initiative that can have a bearing on building resilience to disinformation (Troop, 2017). Another example is Poynter's MediaWise initiative[385], which has delivered MIL face-to-face and online training to over five million teenagers and other citizens, with special focus on under-served communities.

A complementary activity to school-based MIL approaches is **teacher training**, i.e. MIL training aimed at training teachers who can then in turn deliver successful school-based MIL training to students. UNESCO has several resources here and a global process to revise and update a curriculum framework for teachers in the light of recent developments such as the proliferation of disinformation.[386] This is an essential as it enables school-based MIL training to scale up and become sustainable. Examples include the Brazilian Educamidia project[387] and the European YouCheck! project.[388] In France, work with school teachers on the problem of "infox" takes place with the Savoir*Devenir initiative among others.[389]

378 http://project-youcheck.com/game-french/
379 https://factcheckingday.com/lesson-plan
380 https://en.unesco.org/news/media-and-information-literacy-joins-games-learning
381 https://www.africmil.org/programmes-and-projects/media-information-literacy/school-media-clubmedia-in-education/
382 https://mdlab.lau.edu.lb
383 https://lie-detectors.org/
384 https://www.redescordiais.com.br/
385 https://www.poynter.org/mediawise/
386 https://en.unesco.org/sites/default/files/belgrade_recommendations_on_draft_global_standards_for_mil_curricula_guidelines_12_november.pdf
387 https://educamidia.org.br/habilidades
388 http://project-youcheck.com/about/
389 http://savoirdevenir.net/mediatropismes/

Media-content approaches: recent studies show that the older generation (above 50 years old) in some countries has lower than average ability to recognise factual information (Gottfried & Grieco, 2018) and to remember already debunked false claims (Mantzarlis, 2017). One significant challenge is therefore how best to deliver media and information competencies to that target demographic. The previously discussed approaches are not suitable, as older people are much less likely to use online games and rely significantly less on social media platforms as their source of news (Ofcom, 2018a). One promising way is to deliver special programmes through mainstream TV channels. As part of the Beyond Fake News project, the BBC has developed an entire series of documentaries, special reports and features across the BBC's networks in Africa, India, Asia Pacific, Europe, and the Americas which are delivered via TV, radio, and online (BBC, 2018c).

MIL training for online influencers and youth organisations: Digital influencers with their millions of followers have the propensity to spread disinformation widely and thus journalists in Brazil have started dedicated MIL training initiatives aimed to improve the ability of these celebrities to fact-check online content prior to publishing posts in support of false or misleading content (Estarque, 2020). The delivery of MIL training through youth organisations is another promising approach that is being explored with the support of UNESCO in India, Kenya, and Nigeria.[390]

Online verification toolkits and educational materials aimed at improving the general public's understanding of verification and fact checking are also increasingly becoming available e.g. Edutopia[391], a New York Times lesson plan (Schulten & Brown, 2017). UNESCO's MIL CLICKS[392] campaign and its MOOCs promoting media and information literacy, critical thinking, creativity, citizenship, and related skills, with materials in multiple languages (e.g. Arabic, English, French, Greek, Spanish). With support from the UN Office on Drugs and Crime, work was carried out in South Africa on localising a UN model curriculum on disinformation and ethics[393].

In an attempt to make content verification easier to understand, the International Fact Checking Network (IFCN) has produced a 7-step fact checking cartoon[394], currently available in English, French, Italian, Japanese, Portuguese, Serbian, and Swahili. The UK independent fact-checking charity FullFact has produced a similar 10-step misinformation detection toolkit[395], as well as offering a collection of child-oriented literacy materials. Another example is an online educational video "How to spot fake news" by FactCheck.org[396].

Outputs Aimed at Improving Journalistic Professionalism

Firstly, a growing number of **journalist-oriented verification literacy materials** and programmes isbeing created, e.g. the learning module on the history of disinformation (Posetti & Matthews, 2018). UNESCO's Journalism, *Fake News and* Disinformation *Handbook for Journalism Education and Training* (Ireton & Posetti, 2018) is available as a

390 http://www.unesco.org/new/en/communication-and-information/resources/news-and-in-focus-articles/all-news/news/unesco_supported_mil_training_in_india_three_days_of_learni/
391 https://www.edutopia.org/blogs/tag/media-literacy
392 https://en.unesco.org/MILCLICKS
393 http://www.cfms.uct.ac.za/news/media-ethics-workshop-localizes-un-curriculum
394 https://factcheckingday.com/articles/24/this-cartoon-has-7-tips-for-fact-checking-online-information
395 https://fullfact.org/toolkit/
396 https://www.youtube.com/watch?reload=9&v=AkwWcHekMdo&feature=youtu.be

free resource in 11 languages with 30 more translations pending at the time of writing.[397] First Draft also provides courses for journalists in verifying media, websites, visual memes, and manipulated videos[398]. Most recently, they launched a coronavirus-specific resource page too[399]. These are complemented by the latest edition of the "Verification Handbook" (Silverman, 2020), which provides guidance on investigating manipulated content, platforms, and disinformation campaigns.

A second type of shared **resources for journalists** is aimed at **strengthening accuracy in reporting**. Examples include providing a trustworthy resource of curating the latest research on key news topics,[400] current advice on media engagement strategies,[401] a centralised resource of public government data[402] or thoroughly fact-checked information and statistics on the economy, healthcare, immigration, etc.[403] Many of these resources, however, are currently country- and language-specific and are designed for manual human consumption. Their usefulness in fact-checking and content verification can be improved further, if they are also made machine readable/accessible following established data interchange standards.

There is also now awareness that journalists can benefit from the latest academic **research** in the field of disinformation, and even begin to collaborate with researchers, in order to integrate findings from the latest advances in psychology, social science, and data science (Lazer et al., 2017). There is also scope for learning from experts in information influence and strategic communications (Jeangène Vilmer et al., 2018), e.g. around the best debunking strategies for countering disinformation.

As disinformation increases in volume and complexity, journalists increasingly also need help with **learning about newly emerging OSINT[404] (Open Source Intelligence) and content verification tools** and the best practices for their use. Some organisations have now started sharing lists of recommended tools and their usage, e.g. First Draft's tools collection[405], India's BusinessWorld[406]. Widely used specialised tools (e.g. the InVID/WeVerify verification plugin[407]) have also started producing online video tutorials and documentation, to enable journalists to learn the techniques and adopt them quickly in their work. A lack of funding has limited these materials from becoming accessible to journalists in multiple languages.

7.2.4 Response Case Study: COVID-19 Disinformation

In the context of the COVID-19 'disinfodemic', many educational measures are being delivered digitally - often using the same online environments where disinformation proliferates (e.g. social media). These responses are being rolled out especially by MIL

397 https://en.unesco.org/fightfakenews
398 https://firstdraftnews.org/en/education/learn/
399 https://firstdraftnews.org/long-form-article/coronavirus-resources-for-reporters/
400 https://journalistsresource.org/
401 https://mediaengagement.org/
402 https://datausa.io/
403 https://fullfact.org/finder/
404 https://en.wikipedia.org/wiki/Open-source_intelligence
405 https://firstdraftnews.org/long-form-article/coronavirus-tools-and-guides-for-journalists/
406 http://www.businessworld.in/article/5-Tools-Every-Young-Journalist-Should-Learn-About-To-Identify-Fake-News/01-04-2019-16868
407 The InVID/WeVerify verification plugin now offers online tutorials, a virtual classroom, and interactive help: https://weverify.eu/verification-plugin/

projects around the world, media, journalism-oriented civil society organisations and journalism schools, as well as governments.

Examples of media and information literacy projects include:

- Pakistan's *Dawn* newspaper has published a short citizens' guide to surviving the disinfodemic as an act of digital media literacy (Jahangir, 2020).
- The London School of Economics (LSE) has published a guide to helping children navigate COVID-19 disinformation for families forced by the pandemic to homeschool their children (Livingstone, 2020)

Educational interventions aimed at journalists focus on verification, fact-checking and ethical health reporting. Some examples:

- A free online course[408] training journalists how best to cover the pandemic has been developed by the Knight Center for Journalism in the Americas, in partnership with the World Health Organisation (WHO) and UNESCO, with support from the Knight Foundation and the United Nations Development Program (UNDP).
- First Draft's Coronavirus Information Resources page includes a 'debunk database', a curated list of sources, educational webinars about reporting on the pandemic, and tools and guides to aid COVID-19 verification and debunking.
- The African Centre for Media Excellence (ACME) hosts a curated list of resources, tools, tips and sources connected to reporting COVID-19, including a fact-checking collection.
- Afghan NGO NAI has produced "Essentials of journalism performances during COVID 19".[409]
- The Data and Society research group has produced a sheet of 10 tips for journalists covering disinformation.[410]

Of particular importance are cross-border initiatives, such as International Center for Journalists (Barnathan, 2020) with a Global Health Crisis Reporting Forum which includes an interactive, multilingual hub for thousands of journalists around the world. This aims to: aid informed, ethical reporting through direct access to credible sources of scientific and medical expertise; facilitate knowledge sharing and collaborative fact-checking/debunking in reference to COVID-19.

7.2.5 Who are the primary actors behind educational responses and who funds them?

Multi-stakeholder Partnerships: These are MIL initiatives where multiple actors from different categories work together in a partnership. Examples include UNESCO's

408 https://www.ejta.eu/news/free-online-course-journalism-pandemic-covering-covid-19-now-and-future
409 https://nai.org.af/law-and-legal-documents/
410 https://datasociety.net/wp-content/uploads/2020/04/10-Tips-pdf.pdf

MIL global alliance "GAPMIL"[411] and its partnership with Twitter during the annual Media and Information Literacy Week[412], and the AI and media integrity work of the Partnership on AI[413], which comprises over 100 organisations, including all major internet communications companies, some major media organisations, research centres and non-profits. Another example is the MisinfoCon[414] global movement which is specifically concerned with creating tools for verification and fact checking. Its supporters organise tool demos, hackathons, talks, and discussions, including literacy and critical thinking topics.

Civil society organisations and grassroots initiatives: These are MIL programmes and resources created by non-profit organisations and/or citizens. In addition to the examples already discussed above (e.g. First Draft, Drog, LieDetectors), others include the UNESCO-chair supported Savoir*Devenir[415]; the 5Rights foundation with their focus on children data literacy[416]; the Mafindo[417] grassroots Indonesian anti-hoax project; the Google-funded Center for Digital Literacy (CDL)[418] training teacher and school children in Republic of Korea; involvement of youth groups in pan-European MIL projects (e.g. INEDU[419]); grassroots actors producing debunking videos and explainers[420].

Fact-checking Organisations and Networks also provide (mainly journalist-oriented) training sessions and publish training resources, either as individual organisations or through joint initiatives[421]. International fact-checking networks (e.g. the International Fact-Checking Network[422] (IFCN), the First Draft Partner Network[423]) and journalist organisations (e.g. the International Centre for Journalists[424] (ICFJ). Such initiatives frequently attract funding from internet communications companies.

Media organisations are also very active in the development and delivery of MIL, not only through traditional (e.g. TV) and social media channels (e.g. YouTube), but also through direct engagement (e.g. in classrooms or through journalism-oriented training workshops and events). Some examples were discussed in Section 7.2.3 above. Others include the journalist training and education work done by The African Network of Centres for Investigative Reporting (ANCIR[425]) and Code for Africa (CfA[426]);

Government-led initiatives: Many governments have now started running or supporting MIL efforts focused on disinformation. Examples of such initiatives include (many of them

411 https://en.unesco.org/themes/media-and-information-literacy/gapmil/about
412 https://en.unesco.org/news/unesco-partners-twitter-global-media-and-information-literacy-week-2018
413 https://www.partnershiponai.org/ai-and-media-integrity-steering-committee/
414 https://misinfocon.com/join-the-misinfocon-movement-f62172ccb1b
415 http://savoirdevenir.net/chaireunesco/objectifs-missions/
416 https://5rightsfoundation.com/our-work/data-literacy/
417 https://www.mafindo.or.id/about/
418 https://www.blog.google/outreach-initiatives/google-org/digital-and-media-literacy-education-korea/
419 https://in-eduproject.eu/
420 The series is called Smarter EveryDay, ran by the YouTuber engineer Detin Sandlin: https://www.youtube.com/watch?v=1PGm8LslEb4; https://www.youtube.com/watch?v=V-1RhQ1uuQ4; https://www.youtube.com/watch?v=FY_NtO7SIrY
421 Full Fact, Africa Check, and Chequado: https://fullfact.org/blog/2020/feb/joint-research-fight-bad-information/
422 https://www.poynter.org/ifcn/
423 https://firstdraftnews.org/about/
424 https://www.icfj.org/our-work
425 https://investigativecenters.org/
426 https://medium.com/code-for-africa

collated by Poynter[427]) Australia, Belgium, Canada, Denmark, Finland, France, India,[428] Netherlands, Nigeria, Singapore, Sweden, and the U.S.. One example is France's Centre de liaison de l'enseignement et des médias d'information (CLEMI) initiative[429], which perhaps uniquely involves libraries and librarians as key stakeholders in the MIL response. Working at a pan-European level, between 2016 and 2018 the European Union funded 10 projects on MIL and GECD, with more under negotiation from their 2019 funding call. The majority of these were aimed at citizens (e.g. YouCheck!), with the rest targeting journalists and news production (e.g. the The European Media Literacy Toolkit for Newsrooms).

Internet Communication Companies: Educational initiatives undertaken by these companies are aimed at:

- Teaching children MIL skills, e.g. Google's Be Internet Legends[430] and the related YouTube Be Internet Citizens[431] initiatives; Google's global Be Internet Awesome[432] initiative (currently with local resources for Argentina, Belgium, Brazil, Chile, Columbia, Italy, Mexico, Peru, Poland, Saudi Arabia, United Kingdom, and United States);

- Training journalists, improving their technology and skills, and investing in media literacy-oriented editorial projects e.g. the Facebook Journalism Project[433], the Google News Initiative[434] and the related YouTube initiative[435], Google's Fact Check Explorer[436].

7.2.6 How are educational responses evaluated?

Evaluating the success of MIL and GECD initiatives in changing citizen's disinformation consumption and sharing behaviour, is a challenging and largely open problem. According to evidence reviewed by research for this study, it appears that standard metrics and evaluation methodologies are still lacking in maturity. In particular, the challenge is to move beyond the awareness-raising stage, towards sustained and institutionalised MIL interventions that lead to measurable, lasting changes in citizens' online behaviour.

There is also the need for independent evaluation of the impartiality and comprehensiveness of MIL materials and training, in particular those created by the internet communications companies. Concerns have been raised by civil society organisations (5Rights Foundation, 2019) that these tend to focus on making users (especially children but also journalists) focused on false content at the expense of privacy issues, and rather than investing in efforts to fix these problems themselves. Deficiencies

427 https://www.poynter.org/ifcn/anti-misinformation-actions/
428 https://mgiep.unesco.org/
429 http://www.clemi.org/
430 https://beinternetlegends.withgoogle.com/en-gb
431 https://internetcitizens.withyoutube.com
432 https://beinternetawesome.withgoogle.com/en_us
433 https://www.facebook.com/facebookmedia/solutions/facebook-journalism-project
434 https://newsinitiative.withgoogle.com/dnifund/report/european-innovation-supporting-quality-journalism/
435 https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html
436 https://toolbox.google.com/factcheck/explorer

in comprehensiveness and transparency have also been flagged (5Rights Foundation, 2019) with respect to inadequate discussion of the risks arising from algorithmic profiling, automatic content moderation and amplification, and privacy implications of data collection. Similar concerns exist in connection with support from these commercial actors for educational responses designed to strengthen journalism and improve journalists' skills as a response to disinformation.

Further research is needed to study on a large, cross-platform scale, the matter of citizens' exposure and propagation of online disinformation, and on probing the impacts on citizens' understanding and experience of other kinds of disinformation responses. Of particular importance is gauging citizens' knowledge and understanding of the platforms' own algorithmic and curatorial responses and how these impact on disinformation, freedom of expression, right to privacy, and right to information. Due to the recent nature of large-scale online disinformation `wildfires', there is not an extensive body of research with answers to these key questions, and findings are geographically limited and, in some cases, somewhat contradictory. This has motivated policy makers and independent experts to recommend that governments need to invest in further research on these topics (HLEG report, 2018; DCMS report, 2018c; NED, 2018), including not just data science approaches, but also ethnographic studies (NED, 2018).

Researchers have also raised concerns about the recent tendency of focusing MIL primarily on critical thinking about news (Frau-Meigs, 2019). In particular, the concerns are that "they attract funds that could otherwise be attributed to full-fledged MIL projects; they provide one-shot school interventions without much follow-up; they do not scale-up to a national level and reach a limited amount of students."

7.2.7 Challenges and opportunities

An overall challenge is how to help the general public (especially those holding polarised views) to see the value of MIL and invest the time to learn and practice mindful social media engagement behaviour. In addition, MIL faces limits if it does not go wider than news, fact-checking and content verification, without holistic encompassing of wider digital citizenship skills - including freedom of expression and other online and offline freedoms (Frau-Meigs, 2019).

With respect to MIL and GECD responses targeting children, the main challenge is in designing content and delivery mechanisms which are sufficiently engaging and have a lasting impact, as by their very nature, child-oriented responses need to target medium - to long-term outcomes. There is also a challenge to situate MIL initiatives that target disinformation and promote critical thinking within the wider context of online safety education. For instance, there is a need to make the link between the dangers of believing and sharing disinformation on one hand, and the related wider dangers of profiling, micro-targeted advertising, or sharing GPS location on the other.

Similarly, there is a need for appropriate training and education on the professional risks faced by journalists who are frequent targets of disinformation campaigns. These campaigns typically deploy misogyny, racism and bigotry as online abuse tactics designed to discredit and mislead and they require holistic responses that address digital safety and security (Posetti, 2018a).

A target group especially under-served by MIL campaigns is that of older citizens, who according to some research are also more susceptible to believing and spreading

disinformation than other age groups (Vicol, 2020). At the same time, their use, knowledge, and understanding of social platforms can also be quite limited, which adds to the challenge of how best to design and deliver MIL campaigns effectively.

Another challenge that needs to be addressed through educational initiatives is to create awareness of the potentially negative impact of the use of automation in online platforms, namely that automated disinformation moderation techniques employed in some online environments can suffer from algorithmic bias and may discriminate against a specific user group (e.g. girls, racial or ethnic groups). Recent research (5Rights Foundation, 2019) has found that 83% of 11-12 year olds are in favour of platforms automatically removing content by default, without need for it to be flagged by a user. It is unclear however, what proportion of these children are also aware of the freedom of expression implications of unmoderated use of such automation.

A new challenge relates to the COVID-19 crisis. As noted earlier in this study, the World Health Organisation has signalled an "infodemic" meaning an overabundance of content that makes it hard for people to find trustworthy sources and reliable guidance. "[437] In this context of such content overload about the pandemic, the challenge is to develop the capacity of audiences to discern the difference between verified and false or misleading content, as well as to recognise content that is in the process of scientific assessment and validation and thus not yet in either category of true or false. A related challenge is that the educational reactions to the disinfodemic risk being exclusively short-term in focus, losing sight of possible links to long-term and institution-based empowerment programmes and policies to build MIL, including for children and older people, in relation to disinformation in general.

On the opportunity angle, the pandemic has also presented a new focal point for news media and journalists to demonstrate and explain their distinctive role, and a unique moment to sensitise citizens about freedom of expression rights and obligations, provide education to help them, and reinforce MIL and GECD knowledge and skills.

There is also an opportunity that immediate educational initiatives aimed at countering the disinfodemic can be taken with an eye to long-term educational impacts. They can be explicitly structured to ensure lasting MIL outcomes, not only specifically to COVID-19 but also other kinds of health and political or climate disinformation. The crisis provides possibilities for the public to learn to approach most content with scepticism, not cynicism, and to be empowered to make informed judgements about the 'disinfodemic' and the responses to it.

In conclusion, both a massive challenge and a major opportunity that needs to be addressed is that of making MIL and GECD education accessible to children worldwide, estimated to constitute one third of internet users globally and the generation that will in time take charge of informational and other issues (Livingstone et al., 2016). This would require governments around the world to make MIL an integral part of their national school curricula; to invest in professional training of their teachers in MIL; and to work closely with civil society and, media organisations, independent fact-checkers, and the internet communication companies in order to ensure a fully comprehensive, multi-stakeholder media and information literacy provision.

[437] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6

7.2.8 Recommendations for educational responses

The challenges and opportunities identified above, and their significant implications for freedom of expression, provide a basis for the following options for action in this category of responses.

International organisations could:

- Work towards provision of Media and Information Literacy (MIL) educational initiatives and materials aimed at currently under-served countries, languages, and demographics.
- Encourage an holistic approach to MIL that covers freedom of expression issues, as well as disinformation across different topics (such as health, politics and the environment).
- Encourage donors to invest specifically in countermeasures to disinformation that strengthen MIL (as well as freedom of expression, independent journalism, and media development).

Internet communications companies could:

- Integrate MIL into the use of their services, and empower users to understand the full range of issues relevant to disinformation, including fact-checking, algorithmic and labelling issues.
- Foster interdisciplinary action research projects designed to experiment with educational responses to disinformation, and report on these experiments in robust ways that aid knowledge sharing - both across academic disciplines and between industry, educators and researchers.
- Support the development of global and/or regional MIL responses, especially by funding projects in currently under-served regions.[438]

Individual states could:

- Put in place or strengthen MIL policies and resource allocation, especially in the educational system where teachers also need to be trained to deliver MIL to children and youth as a counter to disinformation.
- Earmark funding and support for interventions for older citizens who are both a key voter demographic and a primary risk group for spreading disinformation.
- Support initiatives to address disinformation that targets children, youth, women, unemployed people, refugees and migrants, and rural communities.

News media could:

- Use their platforms to proactively train audiences and internet users about the difference between verified information on the one hand and disinformation on

[438] See, for example: https://www.ictworks.org/african-digital-literacy-project-grant-funding/

the other, and help cultivate the requisite skills to recognise this and navigate the wider content ecosystem, along with the freedom of expression issues involved.

- Support advanced training in verification and counter-disinformation investigative techniques for editorial staff

- Collaborate with journalism schools on counter-disinformation projects involving both researchers and students to improve the capabilities of graduates and deepen their own understanding and practice

Civil society could:

- Increase work in MIL innovation such as anti-disinformation games, and develop creative ways to empower constituencies beyond the educational system who are at risk from disinformation.

- Support the development of global and/or regional MIL responses, especially in currently under-served regions.

- Provide independent evaluation of MIL initiatives carried out and/or supported by internet communications companies.

Researchers could:

- Develop and apply metrics for studying MIL in relation to disinformation.

- Focus on interdisciplinary research to develop new approaches to education as a counter disinformation measure, e.g. integrate methods from journalism studies, computer science, psychology, sociology etc.

- Forge partnerships with news organisations to help strengthen investigative reporting into disinformation and deepen audience insights with reference to engagement with counter-disinformation content.

7.3 Empowerment & credibility labelling responses

Authors: Diana Maynard, Denis Teyssou and Sam Gregory

Educational responses focus on teaching people about the importance of critical thinking and self-awareness in terms of information consumption, thereby giving them internal mental competencies. This chapter looks at empowerment responses that focus specifically on external methods, tools and websites to assist users in the actual understanding of the nature of information and its sources. Thus the two kinds of responses go hand in hand.

As discussed in the previous chapter, teaching media and information literacy to both journalists and citizens alike is one of the significant responses in play. Even if disinformation cannot be wholly thwarted, its dissemination and impact can be reduced if people are able to employ critical thinking in their media and information consumption. This competence underpins the ability to effectively recognise disinformation, along with its appeal and the sources that may promote it. Such awareness can enlist those exposed to falsity to understand their part in preventing its spread and influence.

This chapter complements the educational response focus by examining the efforts around content verification tools and web content indicators that can be seen as aids or prompts that work with people's competencies in the face of disinformation. These tools and cues are intended to help citizens and journalists to avoid falling prey to online disinformation, and to encourage good practices among journalists as well as internet and media companies when publishing information.

This also includes efforts by the news media to boost their credibility over less reliable sources, through highlighting reliable brands and public service broadcasting, as well as methods aimed at consumers for assessing and rating the credibility and reliability of news sources. Examples include Newsguard[439], Decodex[440], the Global Disinformation Index[441], the transparency standards of the Trust Project[442], and a number of browser extensions (many of which are discussed in other chapters of this report, and which are, however, external to consumption of content via apps such as Facebook). Also discussed is the recent emergence of "controlled capture" apps (e.g. TruePic[443]) and newly developed approaches to authentication and provenance tracking that are being considered for use by both individuals and media entities. These include work by the News Provenance Project (NPP)[444], the Content Authenticity Initiative[445], and the complementary Journalism Trust Initiative (JTI)[446].

439 https://www.newsguardtech.com/
440 https://www.lemonde.fr/verification/
441 https://disinformationindex.org/
442 https://thetrustproject.org/
443 https://truepic.com/
444 https://www.newsprovenanceproject.com/
445 https://theblog.adobe.com/content-authenticity-initiative
446 https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation

Bridging from MIL competencies to providing aids to users, are examples like Full Fact's educational toolkit.[447] This resource provides methodologies and suggestions for tools to assist with grasping where news is from, what is missing, and how it makes the reader feel. The initiatives, such as those described here, offer assistance which enables the user to implement these ideas, for example by making it practical to discover what the original source of a piece of information is, and how trustworthy it might be.

Unlike the kinds of fact-checking tools described in Chapter 6.2, which try to prevent the spread of disinformation directly, or which (try to) give a specific answer to the question "is this information true?", empowerment and credibility tools as discussed here instead typically put the onus on the consumer to interpret explicit signals that are given about the content. For example, signals provided by groups such as Credder and Newsguard (described below) provide information about the reliability or credibility of a source, and do not provide answers to whether a specific piece of information is true or not. Similarly, provenance-tracking initiatives show where, and in some cases how, a piece of information originated, but leave it to the user to interpret that (for example, by understanding that if a picture was taken several years ago, it may not be relevant to an event which has just happened, or has been reported as such). There are thus differences to the tools described in the previous section, as will be discussed in more detail in Section 7.3.3.

Evidence about whether something is accurate and credible is often linked to knowing who originally created the content or first shared it. Newsrooms, and people relying on social media for information, need to be investigating the source, almost before they look at the content itself. For example, people should be routinely researching the date and location where this is embedded in it.

As discussed throughout this report, and especially in Chapter 3, disinformation is frequently associated with both domestic and foreign political campaigns, and can lead to widespread mistrust in state authority. One way in which states can both allay the public's fear, and help them to distinguish which information is trustworthy, thereby providing an antidote to some disinformation, is through greater transparency. A strong regime of proactive disclosure by states, along with an effective real time information dispensation, together constitute a buttress to fortify clarity of provenance. However, vigilance must still be maintained because provenance does not equal facticity or comprehensivity. When states do not fully and promptly disclose for example the range of COVID-19 statistics on the channels that are recognisably their own, this is an invitation for understandable rumour and speculation, but also for inauthentic sources to fill the gap with disinformative content.

7.3.1 What and who do empowerment and credibility labelling responses monitor/target?

Provenance-tracking initiatives aim to assist news consumers in understanding the source of information, and thus to be more aware of misleading information, which is complementary to efforts that try to prevent it occurring and spreading in the first place. These initiatives treat attribution information and metadata as tools that can give insight. This is typically relevant to fake images and videos (such as deepfakes, but also ones

[447] https://fullfact.org/toolkit/

which have just been falsely attributed) by means of better authentication. Examples include those by Twitter, Truepic, Serelay, and Amber (further detail below). Alternatively, they may target the ways in which information is displayed to the user, in order to alert them to potentially misleading information such as old content that has (been) resurfaced as if it is current, and sources that might be dubious or untrustworthy. Examples of this include the News Provenance Project and Adobe, as well as Twitter's policy on dealing with manipulated media, which provides in-platform context (Roth & Achuthan, 2020).

Trust-based initiatives, on the other hand, monitor the news providers themselves, attempting to distinguish those which are unreliable, as well as sometimes assessing individual articles and authors. Accreditation-based initiatives largely target the news providers, aiming to "legitimise" those which are most trustworthy. Some, such as Newsguard, also target social media providers and other intermediaries' media, in the hope of financial returns for providing these companies with tools to rank news providers and sources.

7.3.2 Who do empowering and credibility labelling responses try to help?

These initiatives aim at five main types of audience. The majority of them try to help the general public by making them more aware of potential issues, but some also target those who produce or disseminate news, such as journalists and bloggers, as well as the media organisations themselves.

- News consumers are targeted by providing tools which help them to better understand the nature and sources of content (e.g. authentication of, and explicit metadata about, images and videos, better presentation of temporal information such as publishing dates, etc.). This also includes alerting news consumers to media entities who do not meet accepted standards - for example, having a history of suspicious funding, publishing fake material, or other dubious practices.

- News providers are targeted by providing them with methods and tools which can be implemented on their platforms directly, for example through the addition of metadata denoting sources and other credibility-related information for the stories and images/videos they provide, as well as better presentation of credibility-related indicators.

- Journalists are targeted by providing them with tools to help understand the nature of articles and media better (e.g. with provenance and credibility issues).

- Bloggers and citizen journalists, as well as media companies, are targeted by providing good practices and standards which all can follow when producing material (e.g. the Journalism Trust Initiative).

- Internet communications companies are also targeted by tools such as Newsguard. These are seen as a market for services that can help them to recognise purveyors of disinformation, and serve their audiences with these tools.

According to a Pew Research Center study (Mitchell, Gottfried et al., 2019), American news consumers expect the problem of misinformation in the news to be fixed by the news industry, even though they believe it is primarily the politicians who create it. Empowerment and credibility labelling responses put the onus on the consumer (and

sometimes the platform or news media organisations) to filter and interpret the content they encounter. Various research has indicated the potential value of explaining why something might be true/false and providing alternative factual information or a detailed explanation of why information is false (Ecker & Lewandowsky, 2010; Swire & Ecker, 2018). The assumption is that when people have aids for this, in the form of reliability signals, they will be more sceptical in the face of disinformation.

These responses thus aim to facilitate this task for the end user by providing mechanisms to signal disinformation. Changes in the way the news is presented to consumers, for instance, can be used to make the audience more aware of potentially misleading content. In particular, solutions are being proposed for protocols by which informative metadata can be added, made more visible, or even accompany published media wherever it is displayed (e.g. when further shared on social media or in the results of web search), as discussed in the following sections describing such initiatives.

Trust-based initiatives, which focus on highlighting good practices in the media, and promoting and rewarding compliance with professional standards, are based on the idea that particular media sources can be flagged as trustworthy, thereby encouraging common standards and benchmarks which can be adopted by all who produce journalistic content. Ultimately, adopting these standards could pave the way towards processes of certifications. Formal, or even informal, certification could lead to better discriminatory practices by consumers, but also to the adoption of better practices by media producers. An important outcome of trust initiatives is to build the faith of users in the media and counter their fears about the reliability of the content.

The theory of change represented by these initiatives can be summarised as follows:

- Relevant causes of disinformation:
  - news providers or internet communications companies spread disinformation, either because they are not trustworthy themselves and/or because they do not recognise it;
  - users do not recognise it, are influenced by it and also spread it further.
- Actions that the initiatives pursue to address the causes:
  - accrediting trustworthy news sources (and by implication, discrediting untrustworthy ones);
  - developing tools to empower media, internet companies and especially users to make better decisions about which information (and which media sources) can be trusted, as well as signposting issues to journalists and investigators;
  - developing protocols for providing better provenance information and making users aware of the importance of the source of content;
  - developing controlled-capture applications enables creators and distributors of images to create trust in their content.

- Desired outcomes of the initiatives:
  - better discriminatory practices by users;
  - adoption of better practices by news media and internet companies;
  - untrustworthy media sources are called to account;
  - faked media (video, images) become easier to spot and are less easily disseminated;
  - improved understanding by the public of disinformation and its playbook, etc.
- Potential impact of the initiatives:
  - confidence in the media and countering of fears about the reliability of information, leading to improved trust in place of a cynical relativism;
  - reduced rationales for producers of disinformation and encouragement of low-standard media to become more trustworthy;
  - increased spread of accurate information and reduced spread of inaccurate information;
  - increased agency for creators/distributors to assert trustworthiness, and for the users to assess it.

7.3.3 What output do empowerment and credibility labelling responses publish?

These initiatives publish a number of different kinds of output aiming to assist actors, ranging from general information, through methods and protocols, and sometimes even actual tools. These can be summarised as:

- provenance information of source material, and protocols for providing this;
- tools and resources for assessing credibility of news sources, feeding into accreditation schemes and content curation systems;
- methods and protocols for better provision of information to the end user, enabling improved awareness of trustworthy and untrustworthy information and sources;
- tools for rating news sources, articles and authors - carried out either by trained professionals (eg Newsguard) or community-driven (e.g. Credder).

We look at each of these in more detail below.

7.3.3.1 Provenance-tracking initiatives

Provenance-tracking initiatives emanate from a number of sources, and can be divided into three main subgroups: tools from news providers, tools at point-of-capture of images/video, and platform responses.

(**i**) **Tools from news providers** aim to assist news consumers to be more aware of misleading information, rather than try to prevent it occurring in the first place. For example, the News Provenance project aims to help users to better understand the nature and sources of content (e.g. authentication of and explicit metadata about images and videos, better presentation of temporal information such as publishing dates, etc.).

(**ii**) **Tools at point-of-capture of images/video** aim to track enhanced metadata and provenance, and confirm whether images and videos have been altered or not. For example, TruePic[448] is a venture-backed startup which is planning to work with hardware manufacturers (currently, just Qualcomm) to log photos and videos the instant that they are captured. Serelay Trusted Media Capture[449] also enables mobile phones to capture images and videos that are verifiable and for authenticity to be later queried by other apps. Amber[450] produces two tools: Amber Authenticate fingerprints recordings at their source, and tracks their provenance until playback, while Amber Detect uses signal processing and artificial intelligence to identify altered audio and video files. Eyewitness to Atrocities[451] is an app for mobile cameras which was developed for the specific purpose of documenting international crimes such that the footage can be authenticated for use in investigations or trials. Similarly to the others, it automatically records and stores metadata about the time and location of the recording, and includes a traceable chain of custody. All these (and other) tools are discussed in more detail in the Witness report (Witness Media Lab, 2019).

(**iii**) **Platform responses** come directly from the Internet communications companies themselves, e.g. image and video platforms. Some of these encourage users to add information to clarify that content complies with company standards and should not be removed. YouTube highlights the importance of adding context[452], for example to explain why graphic images might be necessary in newsworthy videos (and thus to prevent them being automatically rejected by YouTube in case they get flagged as being dubious). The social video company gives the example of a voice-over narration about the history of a protest - this kind of information is useful in helping a user to understand the provenance of a video.

Other kinds of responses involve directly labelling content; for example, YouTube does this to provide information on videos that highlight conspiracy theories (e.g. around the 1969 Apollo moon landing); or to indicate that content is from a state-funded broadcaster. Other platforms take similar action around inaccurate information on vaccinations, while in August 2020, WhatsApp introduced a feature which signals messages that have been forwarded five times or more, as an indicator of potential viral information.[453] Clicking

448 https://truepic.com/
449 https://www.serelay.com/
450 https://ambervideo.co/
451 https://www.eyewitness.global/
452 https://support.google.com/youtube/answer/6345162?hl=en
453 https://www.theguardian.com/technology/2020/aug/04/whatsapp-launches-factcheck-feature-aimed-at-viral-messages

on the magnifying glass symbol that automatically appears next to such a message initiates an online checking process which aims to reveal any known conspiracy theory or disinformation associated with the content of that message.

In early 2018, YouTube began labelling content in terms of whether the source counted as "state-funded media" in the company's definition.[454] In June 2020 Facebook introduced a similar policy, explaining that it was to help people understand whether the news they read is "coming from a publication that may be under the influence of a government".[455] Twitter introduced the practice some months later.[456]

7.3.3.2 Trust- and accreditation-based initiatives

Trust- and accreditation-based initiatives aim to develop and implement an agreed set of trust and transparency standards for media sources. These standards encompass transparency of media ownership and sources of revenues, as well as journalistic methods and the compliance with ethical norms and independence. Some of them aim to lead to a system of formal accreditation. Examples include:

- the **Journalism Trust Initiative**[457] (which involves Reporters Without Borders and its partners Agence France Presse, and the European Broadcasting Union);
- the **Trust Project**[458] (a consortium of top news companies, including the German news agency dpa, *The Economist*, The Globe and Mail, Hearst Television, the *Independent Journal Review*, Haymarket Media, Institute for Nonprofit News, Italy's *La Repubblica* and *La Stampa*, Reach Plc, and *The Washington Post*, and supported externally by various social media companies and search engines);
- the **Trusted News** initiative[459] set up by the BBC, which is planning a number of collaborative actions such as a rapid-response early warning system so that media (and other) organisations can alert each other rapidly in the case of disinformation which threatens human life. It is particularly tailored towards preventing the disruption of democracy during elections, with other actions based specifically around voter information and media education.

Trust-based initiatives also involve the development of tools and mechanisms for users to rate not only sources, but in some cases also individual articles, and/or journalists in terms of their credibility and trustworthiness. For example, **Credder**[460], which styles itself as "the trusted review site for news media" believes that "news should compete for trust, not clicks". It allows journalists and the public to review articles, measuring trust in not only the articles themselves, but also in the sources cited, and in the authors, and collates statistics on these. More generally, these tools use assessments and scoring of source quality (based on metrics such as accuracy and objectivity) to guide users towards higher-quality information and to help them to better discern and ignore low-quality information.

454 https://money.cnn.com/2018/02/02/media/youtube-state-funded-media-label/index.html
455 https://about.fb.com/news/2020/06/labeling-state-controlled-media/
456 https://www.bbc.co.uk/news/technology-53681021
457 https://jti-rsf.org/
458 https://thetrustproject.org/
459 https://www.bbc.co.uk/mediacentre/latestnews/2019/disinformation
460 http://credder.com

Newsguard[461] offers a browser plugin which aims to rate news sites based on what it defines as good journalism practices, via a nutrition-label methodology. This gives the reader additional context for their news, and also warns advertisers who might be worried about their brand's reputation to avoid unreliable sites. Green-rated sites signal good practices, following basic standards of accuracy and accountability, while red sites signal those with a hidden agenda or which knowingly publish falsehoods or propaganda. Additionally, grey sites refer to Internet platforms, while orange sites indicate satire. A colour-coded icon is shown next to news links on search engines and social media feeds, so that people are informed before they even click on the link. Additional information about the site, such as why it received the rating, can be obtained by hovering the mouse over the icon and/or clicking a button for additional information.

Décodex[462] is a tool created by French newspaper *Le Monde* to help people verify information with respect to rumours, exaggerations, twisted truth, etc. The tool works in two ways: a search tool enabling a user to check the address of a site for more information (e.g. to find out if it is classified as a satirical site); and a browser extension which warns the user when they navigate to a website or a social media account which has been involved in spreading disinformation.

MediaBiasFactCheck[463] is a tool which enables users to check the political bias of a particular media source. In the U.S. in particular, the public's most commonly given reason for media sites not making a clear distinction between fact and fiction is bias, spin and agendas, according to a report by the Reuters Institute (Newman & Fletcher, 2017). According to the philosophy behind the tool, the least biased sites are supposed to be the most credible, with factual reporting and sources provided. Questionable sources, on the other hand, exhibit features such as extreme bias, use of loaded words (conveying strong emotion designed to sway the reader), promotion of propaganda, poor sourcing to credible sites, and a general lack of transparency. However, the methodology behind the approach is not transparent, and it has been criticised itself for its quality.[464] Furthermore, it is not clear that it is a good idea to have a single number telling us how biased a news source is, as the situation is often more complex than this, and any notion of bias requires a baseline.

Maldito Bulo[465] is a browser extension created by one of the main debunking websites in Spain, Maldita.es[466]. The plug-in warns the user who has installed it if the consulted website has already published disinformation and how many stories have been debunked in the domain name.

KnowNews[467] is a browser extension which aims to help users understand which news sites are trustworthy or credible. It is developed by Media Monitoring Africa, which is an independent non-profit organisation from South Africa that promotes media freedom, media quality and ethical journalism.[468] The browser extension automatically classifies news sites based on their credibility, rating sites as credible, "dodgy" or not rated. The tool focuses on the content itself, however, directly evaluating information such as the

461 https://www.newsguardtech.com/
462 https://www.lemonde.fr/verification/
463 https://mediabiasfactcheck.com/
464 https://rationalwiki.org/wiki/Media_Bias/Fact_Check
465 https://chrome.google.com/webstore/detail/maldito-bulo/bpancimhkhejliinianojlkbbajehfdl
466 https://maldita.es/
467 https://newstools.co.za/page/knownews
468 https://mediamonitoringafrica.org/

authenticity of a photo, and is developed in partnership with Facebook and Google, as well as a number of other organisations.

The Knight Foundation's Trust, Media and Democracy Initiative[469] is anchored by the Knight Commission on Trust, Media and Democracy, a panel of people promoting more informed and engaged communities. This non-partisan group provides funding for seven initiatives:

- Cortico data analytics to surface underrepresented voices;
- Duke Tech & Check Cooperative + Share the Facts Database;
- First Draft fact checking network;
- AP fact checking;
- Reynolds Journalism Institute journalist training program;
- Santa Clara University Trust Project trust indicators;
- Your Voice Ohio strengthening ties to local communities;

The **IPTC** (international Press Telecommunications Council - the global standards body of the news media) has been collaborating with several initiatives around trust and misinformation in the news industry since 2018. This mainly involves working with The Trust Project and the Journalism Trust Initiative from Reporters Without Borders, but also to some extent the Credibility Coalition, the Certified Content Coalition and others, with the aim of identifying all known means of expressing trust in news content.

In April 2020, the IPTC published a draft set of guidelines[470] which aim to enable a news agency to add their own trust information to any news items they distribute. These indicators can also be converted to a standard schema.org markup language that can be added to HTML pages and automatically processed by search engines, social media platforms and specialised tools such as the NewsGuard plugin. This then enables users to see the trust indicators and decide for themselves about the trustworthiness of a piece of news.

The aim of the guidelines is to encourage news publishers to use trust indicators to show why they think they can be trusted, rather than just showing a certification of trustworthiness. Readers should be encouraged to follow links to understand the issues better. Indicators include those connected with editorial policy (e.g. statements about disclosure and correction practices, diversity and naming of sources, ethics, and feedback policies); party-level indicators (e.g. lists of other work by the author or provider; awards won; topics of expertise); organisation-level indicators (e.g. staff diversity; founding date of organisation; size etc.); piece-of-work-level indicators (e.g. details about dateline, editor, fact-checking; corrections; provider); person-level indicators (details about the author of the article); and type-of-work indicator (e.g. whether it is satire or not; what kind of report it is; background information, and so on).

469 https://knightfoundation.org/topics/trust-media-and-democracy
470 https://iptc.org/news/public-draft-for-comment-expressing-trust-and-credibility-information-in-iptc-standards/

These guidelines follow the idea of user empowerment by enabling users to make their own decisions rather than following blindly what is suggested to them. It also makes it easier for both information producers and consumers to follow established protocols. However, a limiting factor of this kind of methodology is that the guidelines are quite complex, and it takes time and effort on the part of the user to develop a full understanding about trustworthiness, and then to assess how it matches up to the claims of the organisation or content at hand. This increase in mental effort therefore best suits those who are already of a discerning nature, rather than those most susceptible to disinformation.

Finally, there are some specific themed initiatives which focus on a particular kind of rumour or topic, such as the **Vaccine Confidence Project** (Larson, 2018). This focuses on early detection of rumours about vaccines in an attempt to prevent them gaining impetus, but is entirely manual. A team of international experts monitors news and social media, and also maintains the Vaccine Confidence Index based on tracking public attitudes to relevant issues. While this is primarily a fact-checking operation, the project undertakes related research on trust and risk in this context and is dedicated to building public confidence and mitigating risk in global health. By listening for early signals of public distrust and questioning and providing risk analysis and guidance, they aim to engage the public early and thereby pre-empt potential programme disruptions.

7.3.4 Who are the primary actors behind empowerment and credibility responses and who funds them?

The actors and their sources of funding for these kinds of initiatives are quite varied, ranging from news media, through social media and internet communications companies, through to non-profit monitoring organisations.

Trust and accreditation initiatives are typically funded by either media companies, who are working together to develop formal systems of accreditation, or by monitoring organisations such as Media Monitor Africa and the Journalism Trust Initiative. Media companies clearly have an interest in establishing trust in news sources, though this raises a number of moral dilemmas (see further discussion on this in section 7.3.7 below).

Provenance initiatives are also funded by a variety of sources. Platform responses are typically funded by the social media companies, such as YouTube and Twitter, while other tools are provided by news providers such as the *New York Times*' News Provenance Project. Tools at point-of-capture are often funded by image and video software companies such as Adobe, as well as emanating from dedicated startups such as TruePic[471], Amber[472], and Serelay[473], while the open-source apps typically come from non-profit organisations, e.g. Tella[474], funded by the non-profit organisation Horizontal, and Eyewitness to Atrocities, funded by the International Bar Association in London in partnership with LexisNexis. The Guardian Project, which produces the open-source app ProofMode[475], is funded by a variety of organisations and foundations, including Witness, Google and various governments.

471 https://truepic.com/
472 https://ambervideo.co/
473 https://www.serelay.com/
474 https://hzontal.org/tella/
475 https://guardianproject.info/apps/org.witness.proofmode/

7.3.5 Response Case Study: COVID-19 Disinformation

The COVID-19 pandemic has highlighted the need for ways to help the public become more aware about disinformation. While educational responses are a major type of intervention, with many organisations and governments producing guides to staying well informed about dubious information and rumours surrounding coronavirus, there are also a few specific instances of organisations producing or highlighting credibility labelling and empowerment mechanisms. Many internet communications companies conferred prominent status on sources of reliable information on their services, such as the World Health Organisation and national health ministries. Signposting typically involves providing links to trustworthy sources of information, rather than explicitly pointing to untrustworthy sources. Examples of these efforts include the Harvard Medical School[476], which lists reliable sources of information on corona virus and provides tips on spotting this kind of knowledge resource.

An interesting method for assisting with the flagging of credibility comes from Wikipedia via WikiProjectMedicine[477], a collection of around 35,000 articles monitored by nearly 150 expert editors. Once an article has been flagged as relating to medicine, it becomes scrutinised more closely.[478] In this way, Wikipedia in some sense acts as a role model by having separate standards and stricter rules for particular situations (in this case, public health). Wikipedia also maintains an up-to-date page listing a variety of "misinformation" (as they term it) specifically about the coronavirus.[479]

Complementing these examples are initiatives to flag which content is dubious, thus indirectly also trying to help people to better understand which sources on the pandemic are genuine and provide verified information. (Content deemed dangerous to public and individual health is typically removed when identified). Many guidelines provided typically offer tips to users on not only how to spot a reliable source but also an unreliable ones), and they often offer advice on sharing (as has also been discussed in the first part of this chapter on MILresponses). One initiative to flag disinformation sources around COVID-19 comes from NewsGuard, who have partnered with BT to launch an online toolkit[480] to raise awareness of NewsGuard's online browser plugin, to help the UK public critically assess any information related to the global pandemic they come across online. The initiative is also backed by the UK Department for Culture, Media and Sport (DCMS) and the UK's library association. NewsGuard also made their browser plugin free until the end of July[481], specifically in the light of coronavirus. Previously it was available only as a subscription service, except to users of Microsoft Edge mobile devices. Since 14 May 2020 they have also extended this to all Microsoft Edge users on both mobiles and desktop applications, provided that the extension is used on that brower and downloaded in Microsoft Edge's store. They also set up a Coronavirus Misinformation Tracking Center[482] which signals all the news and information sites in the U.S., the UK, France, Italy, and Germany that they have identified as publishing materially false information about the

476 https://www.health.harvard.edu/blog/be-careful-where-you-get-your-news-about-coronavirus-2020020118801.
477 https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine
478 https://www.wired.com/story/how-wikipedia-prevents-spread-coronavirus-misinformation/
479 https://en.wikipedia.org/wiki/Misinformation_related_to_the_2019%E2%80%9320_coronavirus_pandemic
480 https://www.newsguardtech.com/press/newsguard-partners-with-dcms-and-bt-to-help-counter-spread-of-covid-19-fake-news-as-misinformation-peaks/
481 https://www.zdnet.com/article/newsguard-drops-its-paywall-to-combat-coronavirus-information/
482 https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/

virus.[483] The list includes sites that are notorious for publishing false health content, as well as political sites whose embrace of conspiracy theories extends well beyond politics.

While most credibility labelling responses in the fight against disinformation are still in their infancy due to ethical, legal and technological issues that are not yet fully solved, nevertheless the enormous amount of disinformation around corona virus and its potential seriousness is likely to become a strong driving force towards more effort in developing tools to aid users to supplement their existing media and information literacy levels as they navigate the "infodemic".

7.3.6 How are empowerment and credibility labelling responses evaluated?

Many of these initiatives are highly collaborative in origin and development, and are thus driven and evaluated through community efforts and advisory boards.

- Provenance-based initiatives are largely evaluated in-house. For example, the News Provenance Project is conducting user research to test the effectiveness of their proposed approach, in order to try to discover whether increasing access to metadata and supporting information helps consumers to better understand the veracity of professionally produced photojournalism.

- Accreditation-based initiatives are developing community-based standards with a solid background. For example, the Journalism Trust Initiative involves a very large number of advisory organisations, including official standards bodies such as the French Standardisation Association (AFNOR) and its German equivalent, the German Institute for Standardisation (DIN), as well as the European Committee for Standardisation (CEN).

- Trust-based initiatives are not always formally evaluated, but rely on community-driven input. For example, sites like Credder display the number (and content) of reviews submitted by users, so it is easy to derive statistical information about their usage and about agreement levels. What is less clear, however, is how helpful these reviews are to others. In other words, the quality of the input (the reviews themselves) can easily be judged and collective trust can be assessed, but the usefulness of the output and its overall impact is less easy to understand. Other sites, such as Newsguard, use trained analysts who are experienced journalists, to research online news brands in order to provide the ratings for sites. The lack of formal and independent evaluation for sites could be a major pitfall for such initiatives, especially if quality is dubious and users are unaware (Wilner, 2018: Funke & Mantzarlis, 2018b).

[483] At the end of March 2020, there were 144 sites, though this number is constantly growing.

7.3.7 Challenges and opportunities

There are a number of challenges arising from these kinds of initiatives, in particular the accreditation-based ones. Potential issues around authenticity and labelling are discussed in much fuller detail in the Witness report (Witness Media Lab, 2019), but we give a brief summary below. It should be noted first of all that the current overall impact of these initiatives is quite low since they are not yet widespread, as can be seen by the relatively low number of downloads of many extensions, Even NewsGuard with 78,000 users worldwide is like a drop in the ocean, especially since it is aimed at the general public.

The most important challenges are tackling impartiality, diversity and exclusion, in terms of who makes decisions about credibility and trust, and how they do it. There are also challenges to monitor and renew/revoke accreditation over a time period. Further, as accreditation and trust-based tools are implemented more widely, they become the de facto statements of trust across diverse media environments, resulting in inadvertent exclusion and inclusion of certain media entities. If authentication becomes the default for online media, this has an impact on pluralism as part of freedom of expression, in particular for those who are already disadvantaged in this respect. In particular, it will be a problem for those without access to verification technology or who may not wish to release potentially sensitive information such as their location.[484] This includes those who are in the Global South, use a jailbroken phone[485] , and who are also more likely to be women and live in rural areas. There are also potential problems of weaponisation of authenticity and provenance-based measures where their usage becomes obligatory, particularly in the context of 'fake news' legislation.

Furthermore, if credibility labelling is carried out by companies, there are risks in having a commercial organisation determining what is partisan and what is not, and non-transparent decisions and strategic biases could easily be incorporated. It also risks becoming a "one size fits all" approach, insensitive to cultural and societal specifics of particular countries, and implying that some fact-based narratives are intrinsically more worthy than other fact-based narratives – rather than signalling non-fact based content. Impartiality in such initiatives is often hard to maintain, and determining this on an ongoing basis can be problematic, which also raises questions about periodicity and mechanisms for the continuous review of labelling. Labelling initiatives can also impact negatively on the legitimate diversity of interpretation of narratives that are nevertheless fact-based.

A related issue is the overall accuracy and transparency of labelling tools. The science of provenance and related media forensics is not simple and not easily explained, so that while labelling information sources with their provenance information looks to be a simple solution, it is not only difficult and prone to error, but also not always obvious when it is incorrect. Credibility labelling is also potentially subjective if not opaque to the reader, as already discussed, and also can be error-prone whether manually or automatically carried out. The issues of accuracy around MediaBiasFactCheck have already been discussed earlier in this chapter, and they are certainly not the only tool with debatable quality of results. Who, then, should oversee the quality of such tools?

484 https://www.axios.com/deepfake-authentication-privacy-5fa05902-41eb-40a7-8850-5450bcad0475.html

485 A jailbroken phone is one which has been hacked to free it from the manufacturer's restrictions, and therefore has implications for the software, tracking options, etc. which can be used on it.

Moving on from reliability, we come to the question of psychological responses to the various mechanisms, in particular concerning the idea of empowerment. When content is marked or labelled in some way, there are a number of risks around the idea of interpretation. First, if a large number of false positives are flagged (i.e. if many legitimate or factual pieces of information are indicated to be suspicious or untrustworthy), people become habituated to this and have a tendency to over-interpret false alarms, believing that the tools are just over-sensitive and the labels are not believable. On the other hand, if content is not labelled with provenance or credibility (for example, if it is not clear how it should be labelled, and if false positives are to be avoided), then the assumption might be that the content is trustworthy, which is also potentially dangerous (see Pennycook et al., 2020). In terms of providing users with information, such as explanations around labelling, or even just explanations which help to empower the user in their decision-making, there is a tradeoff between providing sufficient information in enough detail to be clear, and in introducing too much complexity, which perpetuates further the divide between expert and consumer.

More generally, many of these initiatives are still quite young, and there is no broad adoption of any of these credibility labelling, provenance or controlled capture approaches. On the other hand, as discussed above, widespread use of such tools may lead to problems of strategic bias, exclusion, and unintended psychological perceptions. Fundamental questions also arise around how the approaches will be rolled out on a wider scale: for example, whether this will be in collaboration with platforms or in an alternative system. This relates in particular to technologies such as the use of blockchain, and there is a limited application of these approaches outside media and institutions in the Global North.

For content authentication systems at scale, there are issues in how to manage the challenges of doing this across technical and societal implications. For provenance tracking, there are a number of questions around the legitimate privacy and anonymity reasons, such as why people choose not to opt-in, or to only opt in for selected items of content, as well as technical constraints. This leads to the question of how to ensure that trust is on an opt-in rather than an obligation basis, and thereby only a signal of trust, rather than a confirmation. This latter is an important dilemma for many other kinds of anti-disinformation initiatives - in order to be effective, these mechanisms need to be widespread, but this causes serious problems when - sometimes for legitimate privacy reasons such as whistle-blowing - people do not choose to authenticate their data, or when relevant verificatory information is incorrect or missing.

Finally, in terms of user empowerment, there are a number of questions around best practices for managing and presenting complex information. As discussed above, information needs to be presented in a simple yet still meaningful way in order for the general public to be able to make appropriate use of it and understand its implications, but too simple a presentation may lead to misinterpretation by suggesting that issues of verification and trust are black and white. On the other hand, in order not to overwhelm the user, systems of progressive disclosure (by means of breaking down detailed relevant information related to trust and credibility into deeper levels to be explored for further understanding) could be a suitable approach, but have not yet been adopted. Clearly, there is no one-size-fits-all solution to the problem of empowering users to

become more discerning about the information they consume, and further social and psychological research is still very much a requirement, as well as the technical and legal issues to be resolved.

Another challenge linked to the psychological issue is the tradeoff between the empowerment of the user by providing them with pointers to helpful information about provenance and trust, thereby avoiding external bias and simultaneously helping to educate the user, and the fact that the onus is now on the users to make the decisions, when they still may not be sufficiently equipped to interpret the results correctly. The interpretation of labels adds a significant additional neural processing load for the consumer, a known factor in both the spread of disinformation and in issues of unconscious bias and filter bubbles. Conversely, these mechanisms also provide a heuristic shortcut that may not be accurate (see for example a recent report by Witness discussing the history of verified checkmarks, and how they default to erroneous instinctive rather than rational thinking) (Witness Media Lab, 2019). Tools and practices which allow a consumer to verify that a particular piece of content came from a particular source also do not help if the consumer does not properly understand the reliability of that source. Thus a holistic approach that incorporates both aspects, and educates the user to use proper discernment in their news consumption, is still critical.

On the opportunities side, many of the challenges listed above can be addressed through transparency, consultation and respect for pluralism and diversity within freedom of expression. Further, one of the greatest strengths of empowerment and especially credibility labelling responses is that the indicators produced are easy to interpret with little training required. For example, 'traffic light' systems make it very clear what is trustworthy and what is suspicious. This is particularly important for the general public who cannot be expected to become expertly media and information literate overnight, despite the benefits that educational initiatives afford, as discussed in the previous chapter. Nor can the general public be expected to fact-check all the content they come across. Thus, the aids discussed above supplement what skills consumers themselves bring to negotiating with content.

These systems can also lead to long-term benefits such as news providers becoming more trustworthy overall, because when their failings are highlighted compared with certified performers, there is greater incentive to improve. Taking this further, a widespread adoption of sets of certifiable standards for the media industry also has potential benefits, such as helping to strengthen the economic situation of legitimate publishers

Additionally, provenance-tracking initiatives which help the user understand the source and nature of the material, or in some way verify its content, save time. This is important to journalists in the fast-paced media world, but also to ordinary members of the public who do not want to spend a lot of effort in checking sources, even if they understand the importance of it. Along with the trust and credibility tools, this time-saving feature in turn helps to drive the adoption of good practices and standards not only by large media companies, but by all who produce media content, such as bloggers and citizen journalists. Finally, if such solutions are successful, they can be adopted by media organisations more broadly. For example, blockchain-based protocols can be used to share metadata along with media content wherever that content goes.

7.3.8 Recommendations for empowerment and credibility labelling responses

In general, the use of consumer aids entailing standardisation and certification (without compromising pluralism and diversity), as well as approaches that can be rolled out globally and across different platforms, can be encouraged.

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

Internet communications companies and news media could:

- With full respect for media pluralism and diversity, adopt certifiable standards with respect to credibility labelling of news institutions.
- Consider clear and simple, time-saving content labelling approaches, with full transparency about the criteria involved, the implementation process at work, and independent appeal opportunities.
- Avoid quick fix solutions, which can be misleading and have unwanted consequences, such as leading people to blindly trust flags and indicators which may not tell the whole story – or leading to people discounting these signals due to 'false positives' or bias.
- Experiment with signposts and indicators which encourage people to think for themselves, and raise the level of their critical Media and Information Literacy.
- Ensure that empowerment and labelling responses operate in tandem with educational responses for best effect.
- Implement better mechanisms for assuring transparency and accountability of institutions and communities engaged in the design and implementation of empowerment and credibility labelling approaches, as well as their independent evaluation.
- Develop credibility responses with great care, especially with consideration towards less developed countries, smaller media and technology companies, and disadvantaged communities who could be negatively affected by inflexible solutions that are insensitive to inequalities and media pluralism and diversity.

Researchers and civil society could:

- Experiment with the implementation and adoption of global solutions (such as blockchain protocols) for provenance tracking and avoid piecemeal approaches.
- Track practices within the media and internet communications companies as a whole, including assessing the significance of metadata for content no matter where that content ends up.

This study presents an original typology of responses to disinformation which addresses the entire spectrum of responses on a global level, capturing a multitude of initiatives and actors. Moreover, the research offers a unique approach: it places freedom of expression-related challenges and opportunities at the core of the analysis.

Particularly novel, is the study's emphasis on identification and explication of **11 different types of disinformation responses** assessed in terms of the objects they focus upon, instead of framing them through a lens on the key actors involved. Similarly, there is the global scope of the project - with many initiatives included from the developing world to ensure geographical diversity.

Additionally, the diverse nationalities and disciplines of the researchers associated with this project allowed a multiplicity of perspectives to emerge and converge, producing a rich and substantial piece of policy research which is tied to both practice and impact, emphasising technological measures, State interventions, pedagogical initiatives, state and journalistic interventions.

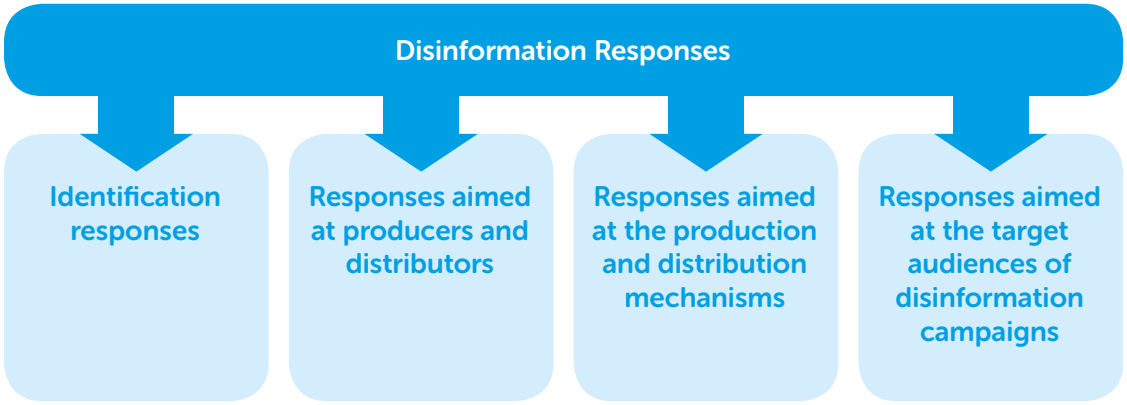
Finally, there is an attempt to deconstruct disinformation in a fresh way, by investigating the underpinnings of these responses in terms of the implied theories of change behind them, as well as an analysis of their targets, and the funding sources they depend upon.

8 Challenges and Recommended Actions

Authors: Julie Posetti and Kalina Bontcheva

8.1 Recapping the typology

At this juncture it is appropriate to summarise the original typology of responses developed by the researchers for this report. Firstly, the types of responses were grouped under **four umbrella categories**:



Then, **11 separate modalities** of response were identified under these four umbrella categories:

1. **Identification responses** (aimed at identifying, debunking, and exposing disinformation)
    - Monitoring and fact-checking
    - Investigative
2. **Responses aimed at producers and distributors through altering the environment that governs and shapes their behaviour** (law and policy responses)
    - Legislative, pre-legislative, and policy responses
    - National and international counter disinformation campaigns
    - Electoral responses
3. **Responses aimed at the production and distribution mechanisms** (pertaining to the policies and practices of institutions mediating content)
    - Curatorial responses
    - Technical and algorithmic responses
    - Economic responses

4. **Responses aimed at the target audiences of disinformation campaigns** (aimed at the potential 'victims' of disinformation)

    - Normative and ethical
    - Educational
    - Empowerment and credibility labelling responses

These different categories of responses are synergistic and symbiotic in nature. They work separately but also interdependently to counter disinformation. For example, normative and ethical responses underpin many of the other response types, while monitoring, fact-checking, and investigative responses play an essential role in informing economic, curatorial and empowerment responses.

8.2 Thematic overview

The landscape mapping and research gap analysis presented in Chapter 3 demonstrates an abundance of studies on disinformation and counter-disinformation methods. However research pre-dating this report largely focuses on the developed West and Anglophone contexts.

In general, while there is a growing body of research, software, training and knowledge resource development focused on tackling disinformation, there is a comparative absence of that which focuses on disinformation in light of human rights, freedom of expression, and the burgeoning access to - and use of - broadband technology worldwide.

At the same time, there is a dearth of research into the impacts of exposure to disinformation and counter-disinformation content on those exposed to it. The research gap analysis also identifies a disconnect between academic research, journalistic investigations, and that commissioned by civil society and intergovernmental organisations. Additionally, it is observed that collaboration between these sectors is infrequent but potentially highly valuable. This report begins to fill the void for research into disinformation responses in the developing world, and it is a starting point for future research endeavours that emphasise linguistic, geographic and disciplinary diversity.

Through an analysis of identification responses (monitoring, fact checking and investigative) to disinformation around the world (Chapters 4.1 and 4.2), the study highlights, in particular, the value of cross-border, interdisciplinary and multi-stakeholder collaboration in initiatives designed to identify falsehoods and fabrications.

Major events, such as elections and public health emergencies, provide an opportunity for independent identification responses to reaffirm the value of facts, and to encourage public reflection of what content they treat as credible and what people decide to share. Identification responses are also important for monitoring the intersection of disinformation with hate speech - against women, minorities, migrants and other vulnerable citizens and communities, including children and the elderly who may be especially susceptible to disinformation.

As discussed, the huge volume, range of sources, and different types of disinformation make it hard to monitor, debunk, assess and raise awareness about the problem. The challenges are many and complex - the importance of working at scale, operating in multiple languages, and across myriad countries in real time being chief among them. The parallel need to achieve impact by limiting and stemming disinformation in measurable ways is also a key challenge, as is the difficulty of evaluating these efforts.

One particular challenge to note is the role of States and political actors as both vectors of, and respondents to, disinformation. However, this study is focused on describing and evaluating disinformation responses and thus the role of such actors as sources and amplifiers of disinformation has not been emphasised. Nevertheless, the prevalence of these sources as instigators and agents of disinformation underscores their obligation to respond to the crisis in transparent and accountable ways.

The importance of defending freedom of expression rights in tandem with responding to disinformation is also underlined through an emphasis on ensuring that fact-checking efforts are neither hampered by partnerships initiated by internet communications companies which may seek to limit the scope of fact-checking or use such initiatives as public relations cover, nor by State and political actors behaving as primary sources of disinformation and key amplifiers of it.

Another issue considered is the failure of verification and fact-checking efforts among some news publishers - either as a product of hyper-partisanship, state capture, 'platform capture', poor standards or under-resourcing - which can turn them into disinformation vectors, possibly eroding audience trust in the process. So, fact-checking and investigative capacity also needs to be developed - both inside journalism, and within civil society organisations that can help reinforce self-regulatory efforts and improve professional accountability.

This represents an opportunity to strengthen identification responses broadly. While some companies and NGOs have pledged some funding to fact-checking organisations, ongoing support throughout, and beyond, critical periods of elections and pandemics is needed. For example, verifying claims about migration and race, vaccinations and climate change will be increasingly important.

At the level of State-based responses to disinformation - including pre-legislative, legislative, policy-based, public education campaigns and electoral responses (chapters 5.1-5.3) - the initiatives are cross-cutting. Actors considered fraudulent or abusive are among the main targets of individual government's regulatory responses, the stated aim being to quell viral incitement to hatred and violence, safeguard against national security risks, deter the disruption of democratic elections, and avoid geopolitical fallout. These responses may also include investing in fact-checking initiatives, public interest journalism and media and information literacy projects. Additionally, they focus on internet communications companies, targeting their economic power and technical behaviour. Based on the assumption that the structure (peer-to-peer distribution) and the algorithms they use enable the amplification of disinformation, many regulatory initiatives attempt to encourage increased responsibility by these actors.

Interventions range from supporting self-regulation through to hard regulatory action that can result in takedown orders and content blocks. Professional disinformation purveyors, such as PR agencies specialising in viral disinformation may also be targeted, along with politicians themselves through new obligations regarding transparency in online political advertising and supporting fact-checking initiatives during elections.

The dramatic pace of technological change renders many attempts at regulation outdated before they are even applied. On the other hand, the associated desire to move quickly to curtail viral disinformation without appropriate debate, transparency and scrutiny entails significant risks for human rights, in particular freedom of expression, press freedom, access to information, and privacy. This can lead to undesirable consequences such as the effective criminalisation of journalism through so-called 'fake news' laws and other measures that may inadvertently or otherwise catch legitimate journalism in the net.

Chapters 6.1 - 6.3 deal with responses aimed at limiting the production and distribution of disinformation: curatorial responses, technical and algorithmic responses, and economic responses. Arguably the biggest challenge identified in connection with these responses is that while recognising the role that internet communications companies need to play in curtailing disinformation published on their platforms, there are potential issues with having (self-)regulatory power delegated to these private companies. This is especially the case where this reduces the accountability and judiciability of expression decisions at large that are the responsibility of States, and which should be in line with international human rights laws and standards. The risk is that censorship can effectively be privatised..

In democratic contexts, such delegation can be explicitly provided by regulations, in which case there can be public accountability for them. However, the companies are largely left to self-regulate content, for a range of political, economic and technological reasons This underlines the urgent need for a robust appeals process and standardised transparency reporting on the way such decisions are taken - to both remove content and to leave it up in the case of targeted abuse and disinformation, for example. This issue is intensifying as the internet communications companies increasingly resort to automation and AI-based algorithms as cost-efficient means for controlling disinformation at scale and at a speed closer to real time. Since algorithms are subject to both potential implicit and explicit bias in their design and in the training data that is used to develop them (with particular implications for gender and racial equality), this is increasingly leading to significant problems, especially in circumstances where the companies have also limited users' ability to resort to a human appeals process. On the positive side, however, at a specific content level, technological responses are less susceptible to external pressure (e.g. from States) applied to individual human operators within a company to take particular action on a case of purported disinformation.

Demonetisation and advertising-linked responses to disinformation are a particular kind of technological response, which are focused specifically on reducing the creation and propagation of disinformation produced for profit. Again, similar to technological responses, the majority of demonetisation responses are largely in the hands of private actors, who are making inconsistent and opaque decisions. In this case, the problems are in the insufficient transparency provided by internet communications companies with regard to advertising, which in effect prevents independent scrutiny and oversight. The problem is acutely present across many platforms and countries not only in the realm of health disinformation (e.g. COVID-19; vaccinations) or issue-based advertisements, but also for political advertising. However, it has become a particular problem for Facebook, which has refused to eliminate micro-targeting of its users in political advertising, and resisted measures to subject direct political speech pertaining to politicians (and their parties and affiliates) to fact-checking, particularly in the United States (Suarez, 2020).

Finally, this study has examined responses aimed at the target audiences of disinformation campaigns: normative and ethical responses, educational responses, and empowerment and credibility labelling responses (chapters 7.1 - 7.3). Responses in the first category - normative and ethical - can be comparatively simple and affordable to implement (although harder to assess in terms of impact), and they can work as counternarratives

that appeal to individuals' moral compasses, or appear to be in alignment with cultural values like anti-racism or anti-misogyny. A problem arises, however, when moral compasses and societal norms are not linked to the principles of access to information, freedom of expression, press freedom, and privacy - as enshrined in international human rights law.

For example, one of the most significant risks associated with State-based responses to disinformation is posed by the kind of legal and regulatory approaches that go against the international norms of freedom of expression (including its corollary press freedom) and privacy. And one of the biggest opportunities exists for internet communications companies to rise above concerns of profit, power and reputation management in order to take genuine action to deal with political disinformation that threatens democracy, in parallel with health disinformation that threatens lives. Ultimately, though, legitimate normative and ethical responses to disinformation can be de-legitimised by the individuals, organisations and States who disagree with the intention behind them, in the same way that credible journalism can be misrepresented as 'fake news' by those seeking to avoid being held to account. That is why these responses need to work in tandem with creative Media and Information Literacy (MIL) interventions, and those that are designed to empower social media users and news consumers.

In the case of educational responses, interventions target a broad spectrum of age groups - from school children to university students, through to journalists, and older citizens, although achieving a comprehensive scale of reach is far from being accomplished in many countries. One of the biggest challenges and opportunities is making Global Citizenship Education (GCED) and Media and Information Literacy (MIL) accessible to children around the world. This is a mission that requires significant investment in interventions (best developed in consultation with multiple stakeholders, including the internet communications companies), and appropriate education for teachers in training as well as for those already working in the field. Another important challenge is to ensure that MIL and GCED interventions are designed holistically and include reference to issues such as hate speech, digital safety and security literacy, as well as identity, human rights and the political economy of media and internet companies' business models. Educational interventions also face the challenge of integrating understanding of the role of journalists and human rights defenders, especially the women and marginalised communities among them, who are frequently targeted in disinformation campaigns. In the bigger picture, it is also important to ensure that a society's emphasis is upon solving the root causes of the disinformation problem, rather than simply building resilience to it as if this was a stand-alone solution.

Lastly, the analysis of empowerment and credibility labelling responses presented here highlighted the important challenges connected with the need to tackle diversity, exclusion and impartiality with reference to who determines what is trustworthy and credible and how it is signalled. There are particular concerns about the implications of these systems for media plurality given the ways in which they can include or exclude media outlets (particularly if they depend upon access to particular technologies or skill sets), and the potential for them to be weaponised in the context of 'fake news' laws. The longer term monitoring, evaluation and updating of these systems was also identified as an area of concern.

Many of the responses to disinformation described in this study are still relatively new and have not yet been broadly adopted. In some cases, this is because certain technologies are still under development or adaptation, because they have not found broad traction, or because there are legal or ethical impediments. For example, when credibility and labelling approaches are not widely used, this clearly impacts on their effectiveness,

and limits the understanding of their potential. This illustrates Collingridge's dilemma (Collingridge, 1980), which essentially posits that the social consequences of technology often cannot be predicted until the technology has already been developed, at which point it is often too late, or at least much more difficult to change. Here, through a convergence of interdisciplinary expertise and pragmatic experience that is focused on a human rights-based assessment of responses to disinformation, an attempt has been made to respond to Collingridge's dilemma at a policy development level.

8.3 Applying the typology to the disinfodemic

In two UNESCO-published policy briefs about responses to COVID-19 disinformation, the typology of responses to disinformation detailed in this report was used to assess the applicability of the formulation to a specific disinformation crisis (Posetti & Bontcheva, 2020a; 2020b). In the context of the COVID-19 disinformation crisis, framed as a 'disinfodemic' by these reports, common and intersecting issues were analysed, and the underlying assumptions, challenges and opportunities of each of the response types were systematically dissected, providing an instructive framework for this more general and comprehensive study.

This process demonstrated the possibility of applying the framework to other specific disinformation crises, such as those associated with elections, anti-vaccination campaigns and climate change.

8.4 Cross-cutting assessment

The responses assessed in this study rest on underlying assumptions, some of which may be open to question and call out for scrutiny. They may be implicit rather than explicit in some cases, and in others they may serve to undermine the intended outcomes of the interventions. Some assumptions may be blind to issues of human rights, while others may incorrectly assume that there would not be unintended effects that harm these rights.

Many of the response modalities presented here seek to strengthen and increase the visibility of genuine public interest information (such as independent journalism, and legitimate public health information, or election safeguarding efforts). Others are aimed at quashing disinformation (or at least downgrading its prominence or pseudo-authoritative character), and there are also instances of responses designed to exert political control or resist regulation.

While the nature of the disinformation problem and its impacts may be assessed differently around the world, and by different actors, all of the interventions presented here are designed to effect change.[486] This is why they have implicit in them a 'theory of change'. What they seek to change and why varies, and the reasons for action may be diverse. For instance, while news organisations fear the impact of disinformation on the value of their journalism (due to online attacks designed to discredit critical reporting, for example) and the business model implications of eroding trust, internet communications companies do not necessarily see disinformation as a problem of economics but rather as a public relations issue and a potential regulatory problem. Governments may wish to regulate for various reasons, one being because they are not satisfied with the companies' responses, or because they see an opportunity to chill critical independent journalism through purported counter-disinformation efforts such as 'fake news' laws. Even though the 'theory of change' behind these interventions is not usually elaborated, the strengths and weaknesses of the particular theory being relied upon are fundamental to the efficacy of the interventions, as well as any unintended effects.

Application of the typology developed as part of this research to specific disinformation crises, such as the COVID-19 'disinfodemic' case study referenced throughout this study, reveals the shortcomings of simply repurposing existing counter-disinformation responses (like those designed to protect elections, for example) to new types and manifestations of disinformation. In the case of COVID-19 disinfodemic responses, pre-existing intervention models (e.g. those applied to climate change denialism and the anti-vaccination movement) were assumed to be an adequate base for responses to the hyper-viral and extremely deadly pandemic-related disinformation. But in many cases, these were not fit for purpose due to the global scale, speed, and range of impacts associated with the pandemic which generated such immense confusion and uncertainty. Far more concerted and complementary interventions, across a wider range of actors, were needed in the face of the 'disinfodemic'.

An underlying assumption in many initiatives rolled out in response to disinformation is that they in effect operate in terms of hunches about what is needed, and how an intervention is expected to work. This is because they operate in the absence of empirical evidence. Few actors dealing directly with disinformation appear to make provision for independent oversight or long term impact assessment, including monitoring and evaluation for unintended effects. Key among these risks is a long-term undermining of the right to freedom of expression, including press freedom, access to information, and privacy protections.

However, a further issue is that accountability for some of the responses is not always obvious or transparent. It is also apparent that many responses are not cognisant of international standards in terms of limitations to freedom of expression rights, in particular with regard to necessity and proportionality. Such overreach infringes the legitimate right to freedom of expression, and especially press freedom which is a precondition for the supply of information that can help overcome the challenge of disinformation.

These accountability issues were exacerbated in the context of the COVID-19 pandemic which demanded swiftly conceived responses rolled out under emergency conditions, in order to deal with an unprecedented global public health threat with massive social and economic ramifications amplified by the disinfodemic.

[486] Note: many of the insights and text that follows have been prepublished in the UNESCO Policy Series "Disinfodemic" https://en.unesco.org/covid19/disinfodemic

8.5 Taking stock of challenges and opportunities

- Time frames: Some responses - like new regulations - are geared towards immediate results, others such as user empowerment are more medium-term. Then, there are measures like developing critical Media and Information Literacy (MIL), which take longer to embed but which may have enduring outcomes. Others - like support measures for journalistic coverage designed to counter disinformation - are more time-specific. It is worth noting that different problems and opportunities operate within different time-frames.

- Complementarities: The 11 types of responses to disinformation modeled here are in many ways complementary to each other. They can be recognised as a holistic package of interventions. For example, in many cases, journalists have exposed online disinformation that had remained undetected (or unrecognised) by the internet communication companies enabling its transmission. In the bigger picture of responses, actions by these companies need to receive attention. This is because the use of power and policy, and the attention to audiences, are the categories of responses that cannot alone 'fix' the disinformation problem - they need to work in tandem with actions taken by the industry to stop transmission of disinformation.

- Contradictions: There are cases where one type of response can work against another. An example would be an imbalance whereby there is over-emphasis on having top-down regulation, while at the same time neglecting the need for bottom-up empowerment. Another example would be the resistance of internet communications companies to removing content associated with disinformation-laden attacks on journalists on the grounds of 'free speech'. This highlights a tension whereby 'free speech' can be justified as a reason to avoid responsibility for responding swiftly and decisively to disinformation that actively undermines press freedom (a corollary of freedom of expression) and journalism safety. In other words, preserving 'free speech' without observing and preserving press freedom rights (which include the need to protect journalists) is not a sustainable approach to mitigating overreach in responses to disinformation.

- Another tension would be the act of catching journalists in nets set for disinformation agents through the criminalisation of the publication or distribution of false information (e.g. via 'fake news' laws), precisely when journalism is needed to counter disinformation. It can also be noted that counter-disinformation content needs to coexist with, not compete with, nor be at the expense of, independent journalism. The different interventions therefore need to be aligned, rather than going in separate directions.

- Gender: There is gender-blindness in many of the responses to disinformation, which risks missing the subtle differences in how false content often targets people, as well as overlooking differences in the way people respond to the content concerned. It is also important to note that established patterns of behaviour by disinformation agents include gendered attacks online (ranging from abuse and threats of sexual violence to digital security and privacy breaches). There is also the issue of women and girls' access to information, which is often

restricted in certain contexts, and threatened by the presence of domestic violence, potentially limiting their access to counter-disinformation efforts.

- Age demographics, particularly regarding children and older people in response to the disinfodemic are also under-considered in many of the responses.

8.6 Overview assessment

Disinformation thrives in the absence of verifiable, trustworthy information. Equally, it can also flourish amid high volumes of content when people may find it difficult to distinguish credible information from disinformation, between what is a verified fact and what is not. It exploits people's need for sense-making of complex developments, as well as their fears, hopes and identities. This is why a multi-faceted approach is needed - one that also goes beyond the realm of communications and contested content, to include practical steps like social inclusion and solidarity, the reinforcement of ethics and values at the personal and community levels, and the embedding of peace-building principles within online communities. Any coherent strategy to fight the realm of information pollution also needs to recognise the value of securing a holistic and analytical approach to the problem.

In this wider context, it is evident that freedom of expression, access to information and independent journalism - supported by open and affordable internet access - are not only fundamental human rights, but also essential parts of the arsenal against chronic disinformation - whether connected to a pandemic, elections or climate change.

It should be noted that the fight against disinformation is not a call to suppress the pluralism of information and opinion, nor to suppress vibrant policy debate. It is a fight for facts, because without evidence-based information for every person, access to reliable, credible, independently verifiable information that supports democracy and helps avert worsening the impacts of crises like pandemics will not be possible.

8.7 Disinformation Responses: Freedom of Expression Assessment Framework

This 23-step assessment tool is designed to assist UNESCO Member States to formulate legislative, regulatory and policy responses to counter disinformation at the same time as respecting freedom of expression, access to information and privacy rights. The tool could be applied to proposed legislation and policy in development to assess - step by step - appropriateness in reference to international human rights laws and norms.

1. Have responses been the subject of multi-stakeholder engagement and input (especially with civil society organisations, specialist researchers, and press freedom experts) prior to formulation and implementation? In the case of legislative responses, has there been appropriate opportunity for deliberation prior to adoption, and can there be independent review?

2. Do the responses clearly and transparently identify the specific problems to be addressed (such as individual recklessness or fraudulent activity; the functioning of internet communications companies and media organisations; practices by officials or foreign actors that impact negatively on e.g. public health and safety, electoral integrity and climate change mitigation, etc.)?

3. Do responses include an impact assessment as regards consequences for international human rights frameworks that support freedom of expression, press freedom, access to information or privacy?

4. Do the responses impinge on or limit freedom of expression, privacy and access to information rights? If so, and the circumstances triggering the response are considered appropriate for such intervention (e.g. the COVID-19 pandemic), is the interference with such rights narrowly-defined, necessary, proportionate and time limited?

5. Does a given response restrict or risk acts of journalism such as reporting, publishing, and confidentiality of source communications, and does it limit the right of access to public interest information? Responses in this category could include: 'fake news' laws; restrictions on freedom of movement and access to information in general, and as applied to a given topic (e.g. health statistics, public expenditures); communications interception and targeted or mass surveillance; data retention and handover. If these measures do impinge on these journalistic functions or on accountability of duty-bearers to rights-holders in general, refer to point 4. above.

6. If a given response does limit any of the rights outlined in 4., does it provide exemptions for acts of journalism?

7. Are responses (e.g. educational, normative, legal, etc.) considered together and holistically in terms of their different roles, complementarities and possible contradictions?

8. Are responses primarily restrictive (e.g. legal limits on electoral disinformation), or there is an appropriate balance with enabling and empowering measures (e.g. increased voter education and Media and Information Literacy)?

9. While the impact of disinformation and misinformation can be equally serious, do the responses recognise the difference in motivation between those actors involved in deliberate falsehood (disinformation) and those implicated in unwitting falsehood (misinformation), and are actions tailored accordingly?

10. Do the responses conflate or equate disinformation content with hate speech content (even though international standards justify strong interventions to limit the latter, while falsehoods are not per se excluded from freedom of expression)?

11. Are journalists, political actors and human rights defenders able to receive effective judicial protection from disinformation and/or hateful content which incites hostility, violence and discrimination, and is aimed at intimidating them?

12. Do legal responses come with guidance and training for implementation by law enforcement, prosecutors and judges, concerning the need to protect the core right of freedom of expression and the implications of restricting this right?

13. Is the response able to be transparently assessed, and is there a process to systematically monitor and evaluate the freedom of expression impacts?

14. Are the responses the subject of oversight and accountability measures, including review and accountability systems (such as reports to the public, parliamentarians, specific stakeholders)?

15. Is a given response able to be appealed or rolled-back if it is found that any benefits are outweighed by negative impacts on freedom of expression, access to information and privacy rights (which are themselves antidotes to disinformation)?

16. Are measures relating to internet communications companies developed with due regard to multi-stakeholder engagement and in the interests of promoting transparency and accountability, while avoiding privatisation of censorship?

17. Is there assessment (informed by expert advice) of both the potential and the limits of technological responses which deal with disinformation (while keeping freedom of expression and privacy intact)? Are there unrealistic expectations concerning the role of technology?

18. Are civil society actors (including NGOs, researchers, and the news media) engaged as autonomous partners in regard to combatting disinformation?

19. Do responses support the production, supply and circulation of information - including local and multilingual information - as a credible alternative to disinformation? Examples could be subsidies for investigative journalism into disinformation, support for community radio and minority-language media.

20. Do the responses include support for institutions (e.g. public service messaging and announcements; schools) to enable counter-disinformation work? This could include interventions such as investment in projects and programmes specifically designed to help 'inoculate' broad communities against disinformation through media and information literacy programmes.

21. Do the responses maximise the openness and availability of data held by state authorities, with due regard to personal privacy protections, as part of the right to information and official action aimed at pre-empting rumour and enabling research and reportage that is rooted in facts?

22. Are the responses gender-sensitive and mindful of particular vulnerabilities (e.g. youth, the elderly) relevant to disinformation exposure, distribution and impacts?

23. If the response measures are introduced to respond to an urgent problem, or designed for short term impact (e.g. time sensitive interventions connected to elections) are they accompanied by initiatives, programmes or campaigns designed to effect and embed change in the medium to long term?

8.8 Comprehensive recommendations for action

The recommendations below build upon the chapter-specific recommendations on particular types of disinformation. They aggregate key points from the chapters in order to set out a full list of options for each individual stakeholder group in regard to the range of disinformation types. This gives an easy-to-use overview of the holistic range of actions which each stakeholder group can consider undertaking in order to optimise the effectiveness and freedom of expression dimensions of their responses. At the same time, partnerships within and across each stakeholder group are recognised as essential for success.

Cross-cutting recommendations aimed at all actors:

- Encourage the strengthening of the range of diverse responses to disinformation, and ensure that these are all in line with international human rights standards.
- Facilitate and encourage coordinated, global multi-stakeholder cooperation and exchange of good practice across continents and states, towards effective implementation of holistic measures for tackling online disinformation.
- Encourage donors to invest specifically in countermeasures to disinformation that strengthen Media and Information Literacy, freedom of expression, independent journalism and media development.
- Increase official transparency and proactive disclosure of official information and data, and monitor this performance in line with the right to information and SDG indicator 16.10.2 that assesses the adoption and implementation of constitutional, statutory and/or policy guarantees for public access to information.
- Promote privacy-preserving, equitable access to key data from internet communications companies, to enable independent analysis into the incidence, spread and impact of online disinformation on citizens around the word, and especially in the context of elections, public health, and natural disasters.
- Invest in independent research into the fast-moving nature and scale of disinformation responses, as well as the need to address the challenges of studying new and rapidly evolving social platforms, including those received or perceived mainly as entertainment and social spaces (e.g. TikTok).

The Broadband Commission could:

- Continue monitoring, measuring and assessing the impacts of responses to disinformation against human rights frameworks, including use of the assessment framework presented above.
- Encourage members who are internet communications companies to ensure the responses that they initiate are appropriately transparent and measurable, as well as implemented on a truly global scale.

- Encourage member companies to consider swift and decisive responses to political and electoral disinformation, as has happened in the field of COVID-19 related disinformation, with due regard to the difference between these two subject fields.

- Encourage its members to integrate this study into their activities, and to bring it to the attention of their stakeholders.

Intergovernmental and other international organisations, **as appropriate**, **could**:

- Increase technical assistance to Member States at their request in order to help develop regulatory frameworks and policies, in line with international freedom of expression and privacy standards, to address online disinformation. This could involve encouraging the uptake of the 23-step disinformation response assessment framework developed for this study.

- Particularly in the case of UNESCO with its mandate on freedom of expression, step up the work being done on disinformation in partnership with other UN organisations and the range of actors engaged in this space.

- Invest in researching, monitoring, measuring and assessing the impacts of responses to disinformation against human rights frameworks, including using the assessment framework presented here.

- Work together with States and NGOs towards Media and Information Literacy initiatives targeting potentially-vulnerable groups.

- Consider convening multilingual conferences, knowledge sharing, and workshops focused on Media and Information Literacy as a response to disinformation.

- Increase work in Media and Information Literacy and training of journalists as significant responses to disinformation.

- Increase support to media institutions in developing countries, including through UNESCO's International Programme for the Development of Communications (IPDC) to enable them to continue producing public interest journalism, and equipping them to combat disinformation.

- Support gender sensitive responses to disinformation.

- Encourage donors to invest specifically in countermeasures to disinformation that strengthen independent fact checking, Media and Information Literacy, freedom of expression, independent journalism and media development.

Individual states could:

- Actively reject the practice of disinformation peddling, including making a commitment not to engage in public opinion manipulation either directly or indirectly - for example, via 'influence operations' produced by third party operators such as 'dark propaganda' public relations (PR) firms.

- Review and adapt their responses to disinformation, using the 23-step framework for assessing law and policy developed as an output of this study, with a view to conformity with international human rights standards (notably freedom of

expression, including access to information, and privacy rights), and at the same time making provision for monitoring and evaluation of their responses.

- Increase transparency and proactive disclosure of official information and data, and monitor this performance in line with the right to information and SDG indicator 16.10.2 that assesses the adoption and implementation of constitutional, statutory and/or policy guarantees for public access to information.

- Promote affordable connectivity for all in line with UNESCO's concept of 'Internet Universality and the four ROAM principles (Rights, Openness, Accessibility and Multi-stakeholder participation).

- Support transparent social media councils and/or national ombuds facilities in order to help give users recourse to independent arbitration and appeals against moderation steps perceived to be unfair.

- Support investment in strengthening independent media, including community and public service media, as the economic impacts of the COVID-19 crisis threaten journalistic sustainability around the world.

- Earmark funding and support for Media and Information Literacy focused on combatting disinformation, especially through educational interventions targeting children, young people, older citizens, and potentially-vulnerable groups.

- Work with internet communications companies to establish privacy-preserving, secure data exchanges and facilitate access to social media data for journalists, and academic researchers, and NGO-based researchers where appropriate, to enable thorough investigations and preservation of historically-important data (especially as associated with elections, pandemics and other important flashpoints).

- Avoid criminalising disinformation to ensure legitimate journalism and other public interest information is not caught in the nets of 'fake news' laws, etc.

- Ensure that any legislation or regulation responding to disinformation crises, like the COVID-19 'disinfodemic', is necessary, proportionate and time-limited.

- Develop mechanisms for independent oversight and evaluation of the freedom of expression implications and efficacy of 'fake news' legislation, along with other relevant national policies and normative initiatives.

- Ensure gender sensitivity in their strategies and public responses to disinformation

- Encourage the uptake of the recommendations below for political parties and actors in reference to elections and campaigning.

Electoral regulatory bodies and national authorities could:

- Strengthen legal measures concerning privacy protection, freedom of expression and political advertising in order to better protect against electoral disinformation.

- Improve transparency of election advertising by political parties, candidates, and affiliated organisations through requiring comprehensive and openly available ad databases and disclosure of spending by political parties and support groups.

- Establish effective cooperation with internet communication companies on monitoring and addressing threats to election integrity.
- Seek to establish and promote multi-stakeholder responses including especially civil society.
- Educate and empower citizens to detect and report disinformation during elections.
- Develop voter literacy through linking civics literacy with digital citizenship education and Media and Information Literacy.
- Work with journalists and researchers in fact-checking and investigations around electoral disinformation networks and producers of 'dark propaganda'.

Political parties and other political actors could:

- Speak out about the dangers of political actors as sources and amplifiers of disinformation and work to improve the quality of the information ecosystem and increase trust in democratic institutions.
- Refrain from using disinformation tactics in political campaigning, including the use of covert tools of public opinion manipulation and 'dark propaganda' PR firms.
- Consider following in the footsteps of political parties in recent elections where the contestants pledged to avoid disinformation.[487]
- Commit to transparency and accountability regarding scrutiny by critical journalistic actors and other mechanisms supporting open societies, and condemn threats against journalists including the use of disinformation as a weapon against the news media.
- Submit their online political adverts to independent fact-checking processes.

Law enforcement agencies and the judiciary could:

- Ensure that law enforcement officers are aware of freedom of expression and privacy rights, including protections afforded to journalists who publish verifiable information in the public interest, and avoid arbitrary actions in connection with any laws criminalising disinformation.
- For judges and other judicial actors: Pay special attention when reviewing laws and cases related to addressing measures to fight disinformation, such as criminalisation, in order to help guarantee that international standards on freedom of expression and privacy are fully respected within those measures.

[487] In Uruguay, political parties in 2019 agreed a pact to refrain from disinformation; to avoid actions or expressions that use aggravating tones against adversaries; and to set up a consultation mechanism when threats or challenges arise to fulfilment of their agreement. https://www.undp.org/content/dam/uruguay/docs/GD/undp-uy-pacto-etico-definformacion.pdf/ In Germany, political parties committed to avoiding social media 'bots' and microtargeting. https://www.bpb.de/gesellschaft/digitales/digitale-desinformation/290568/relevanz-und-regulierung-von-social-bots

Internet communications companies could:

- Intensify multi-stakeholder engagement and transparency about their policies in general and application thereof, including their responses to disinformation.
- Implement their responses on a global scale, rather than being limited to certain countries, and ensure coverage in all significant languages.
- Provide more financial support to: independent fact-checking networks, independent journalism (especially those focused on investigations targeting disinformation content and networks, and also to local news organisations which are particularly fragile), and independently-provided/delivered Media and Information Literacy initiatives.
- Avoid interventions that appear designed primarily as public relations or brand management exercises, make contributions with 'no strings attached', and improve transparency related to such funding.
- Support independently managed funds for research into cases of disinformation, its impact, and responses to it, including independent evaluations of the effectiveness of companies' own disinformation responses. Ensure a diversity of funding recipients along with transparency regarding the research methods and findings.
- Work together, through a human rights frame, to deal with cross-platform disinformation, in order to improve technological abilities to detect and curtail problems of false and misleading content more effectively, and share data about this.
- Develop curatorial responses to ensure that users can easily access journalism as verifiable information shared in the public interest, prioritising news organisations that practice critical, ethical independent journalism.
- Work to boost the visibility of credible news content and financially compensate news producers whose content benefits their businesses, especially as many news organisations removed paywalls and other barriers to content access during the COVID-19 pandemic as a counter-disinformation measure.
- Avoid overreliance on automation for content moderation, recognise the need to expand human review capacity and remedies for redress, and transparently monitor these matters.
- Ensure appropriate pay, training and psychological support for the people working in content moderation.
- Recognise that if health disinformation and misinformation can be quickly dealt with in a pandemic on the basis that it poses a serious risk to public health, action is also needed against political disinformation - especially at the intersection of hate speech – when it too can be life-threatening. The same applies to disinformation related to climate change.
- Recognise that press freedom and journalism safety are critical components of the right of freedom of expression, meaning that online violence targeting journalists (a frequent feature of disinformation campaigns) cannot be tolerated.

- Apply fact-checking to all political content (including advertising, fact-based opinion and direct speech) published by politicians, political parties, their affiliates and other political actors.

- Produce detailed and frequent public transparency reports, including specific information on identification of the origins, scale, views, flow and types of disinformation, removals of disinformation, demonetisation of disinformation content, and suspension of accounts spreading disinformation, as well as provide information on other curational steps such as labelling and appeals.

The media sector could:

- Redouble their efforts as professional frontline responders to disinformation, through increased investment in fact-checking, debunking, disinformation investigations, and ensuring robust lines of questioning about responses to disinformation, as well as by enhancing accountability and transparency with regard to political actors, states, institutions, and the corporate sector.

- Report on the human rights implications of responses to disinformation, including those impacting on freedom of expression and access to information, as well as privacy rights.

- Consider mythbusting and investigative collaborations into disinformation with other news organisations and audiences, including internationally. Partnerships with researchers and civil society organisations can also be successful.

- Focus innovation efforts on countering disinformation through accessible and engaging story formats, such as infographics and podcasts along with collaborative, data-driven investigations.

- Ensure that experiences in a range of developing countries are not overlooked in coverage of disinformation and responses to it.

- Ensure preparedness of staff for safety risks associated with reporting on disinformation, e.g. increased security threats, online abuse, physical attacks, and ensure gender sensitivity in responding to these dangers.

- Undertake coverage of the issues of transparency, accountability and independence of institutions and individuals engaged in fact-checking and/or evaluation of the credibility of sources of information.

Civil society could:

- Reinforce the call for responses to disinformation to conform to international human rights standards.

- Partner with journalists, news organisations and researchers on investigative and monitoring projects about disinformation and responses to it.

- Strengthen the roll-out of Media and Information Literacy projects, and of programmes that support independent journalism.

- Consider programmes targeting children as well as older citizens who are under-served by Media and Information Literacy campaigns, and therefore more susceptible to exploitation by disinformation agents.

- Produce counter-content and campaign against disinformation.

Researchers could:

- Strengthen their scientific enquiry agendas to focus on disinformation, the responses to it, and the impacts of these responses.

- Study under-researched formats such as interactive gaming where disinformation and countermeasures may effectively target young people.

- Undertake Participatory Action Research projects that respond to critical incidents connected to disinformation, and can also provide urgent knowledge.

- Collaborate with journalists, news organisations, and civil society groups on projects that help surface and combat disinformation, along with monitoring and assessment exercises focused on responses to it.

- Study cross-platform disinformation campaigns to get a more rounded, holistic perspective on the problem and responses to it.

- Pursue independent, longitudinal, quantitative and qualitative monitoring and evaluation of disinformation responses implemented by the internet communications companies.

- Develop new technological tools to assist journalists and other verification professionals in detecting and analysing disinformation, ensuring also that such tools put freedom-of-expression, privacy, algorithmic transparency and accessibility at their core.

- Use the typology and assessment framework developed through this study to analyse various types of disinformation responses as they emerge and evolve, and assess their efficacy and impacts with specific reference to freedom of expression challenges.

- Invest in studies that address the impacts of disinformation on consumers, including behavioural science investigations that build knowledge about what motivates people to share and/or give credence to disinformation.

- Prioritise studies targeting users' behaviour in relation to engagement with, and redistribution of, credible, verified information such as that produced by independent news publishers and journalists.

- Ensure female experts are visible as a way of addressing gender inequalities in international debates on disinformation.

List of Sources Consulted

9

Aaron, C. (2020). 'Journalism Needs a Stimulus. Here's What it Should Look Like'. *CJR*, 24 March 2020. https://www.cjr.org/analysis/journalism-stimulus.php

Abellán, L. (2019). 'El Gobierno activa una unidad contra la desinformación ante las elecciones'. *El Pais*, 11 March 2019. https://elpais.com/politica/2019/03/10/actualidad/1552243571_703630.html

Adair, B. (2018). 'Beyond the Truth-O-Meter'. *Columbia Journalism Review*, 24 July 2018. https://www.cjr.org/first_person/beyond-the-truth-o-meter.php

Adobe. (2019). 'Introducing the Content Authenticity Initiative'. 04 November 2019. https://theblog.adobe.com/content-authenticity-initiative/

Adhikari, R. (2020). ''Black PR' Firms Line Their Pockets by Spreading Misinformation'. *Ecommerce Times*, 09 January 2020. https://www.ecommercetimes.com/story/86444.html

AEC. (Australian Electoral Commission). (2019a). 'Electoral Integrity: 2019 Federal Election'. https://www.aec.gov.au/elections/electoral-advertising/electoral-integrity.htm

AEC. (Australian Electoral Commission). (2019b). 'Stop and Consider'. https://www.aec.gov.au/Elections/electoral-advertising/stopandconsider.htm

AFP. (2019a). 'Sri Lanka proposes new law on fake news after Easter attacks'. *France 24*, 05 June 2019. https://www.france24.com/en/20190605-sri-lanka-proposes-new-law-fake-news-after-easter-attacks

AFP. (2019b). 'Nine arrested in Thailand for posting election 'fake news'. *France 24*, 28 March 2019. https://www.france24.com/en/20190328-nine-arrested-thailand-posting-election-fake-news

AFP. (2019c). 'Côte d'Ivoire: le député pro-Soro Lobognon condamné à un an de prison pour un tweet'. L*a Libre Afrique*, 30 January 2019. https://afrique.lalibre.be/31666/cote-divoire-le-depute-pro-soro-lobognon-condamne-a-un-an-de-prison-pour-un-tweet/

AFP. (2020). 'Legal challenge to Singapore misinformation law rejected'. 05 February 2020. http://theindependent.sg/legal-challenge-to-singapore-misinformation-law-rejected/

Africa Check, Chequeado, Full Fact. (2020). 'What is the impact of fact checkers' work. 13 March 2020. https://fullfact.org/blog/2020/mar/long-game-impact-fact-checkers/

Agrawal, A. (2019). 'Intermediary Guidelines to be notified by Jan 15, 2020, MeitY tells Supreme Court', *Medianama*, 21 October 2019. https://www.medianama.com/2019/10/223-intermediary-guidelines-to-be-notified-by-jan-15-2020-meity-tells-supreme-court/

Agarwal, S, Farid, H., Gu, Y., He, M., Nagano, K. & Li, H. (2019). 'Protecting World Leaders Against Deep Fakes'. *CVPR Workshop Paper*. http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf

Ainge Roy, E. (2019). 'New Zealand bans foreign political donations amid interference concerns'. *The Guardian*, 03 December 2019. https://www.theguardian.com/world/2019/dec/03/new-zealand-bans-foreign-political-donations-amid-interference-concerns

Alam, J. (2018). 'Journalists Slam Pending Bangladesh Digital Security Law', *Associated Press*, 28 September 2018. https://apnews.com/6c7914d0733249aabacd50a3c40542fe/Journalists-slam-pending-Bangladesh-digital-security-law

Alaphilippe, A., De Marez, L., Gerlache, A., Lievens, E., Pauwels, T., Picone, I. & Rouvroy, A. (2018a) 'Verslag van de Belgische Expertengroep inzake Fake News en Desinformatie'. 18 July 2018. https://www.dropbox.com/s/99iza9kmbwjbels/20180718_rapport_onlinedesinformatieNL.pdf?dl=0

Alaphilippe, A., Bontcheva, K., Gizikis, A. Hanot, C., (2018b). 'Automated tackling of disinformation'. *European Parliamentary Research Service. Scientific Foresight Unit* (*STOA*). PE 624.278, March 2019. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf

AlBadawy, E. A. Lyu, S. & Farid, H. (2019). 'Detecting AI-Synthesized Speech Using Bispectral Analysis'. *ResearchGate*, June 2019. https://www.researchgate.net/publication/333393640_Detecting_AI-Synthesized_Speech_Using_Bispectral_Analysis

Al Busaidi, A. S. H. (2019). 'Tackling fake news head-on in Oman'. *Times of Oman*, 28 July 2019. https://timesofoman.com/article/1681323/Oman/Tackling-fake-news-head-on-in-Oman

Alexander, J. (2020). 'YouTube is demonetizing videos about coronavirus, and creators are mad'. *The Verge*, 04 March 2020. https://www.theverge.com/2020/3/4/21164553/youtube-coronavirus-demonetization-sensitive-subjects-advertising-guidelines-revenue

Al Jazeera. (2018). 'Bangladesh Shuts Down Mobile Internet in Lead Up to Election Day'. 28 December 2018. https://www.aljazeera.com/news/2018/12/bangladesh-shuts-mobile-internet-lead-election-day-181229111353218.html

Allan, R. (2018). 'Hard Questions: Where Do We Draw the Line on Free Expression?' *Facebook*, 09 August 2018. https://about.fb.com/news/2018/08/hard-questions-free-expression/

Allcott, H, Gentzkow, M. (2017). 'Social Media and Fake News in the 2016 Election'. *Journal of Economic Perspectives* - Volume 31, Number 2, Spring 2017.

Allcott, H. Gentzkow, M. & Yu, C. (2018). 'Trends in the Diffusion of Misinformation on Social Media'. *Stanford, Institute for Economic Policy Research* (*SIEPR*). https://siepr.stanford.edu/system/files/publications/18-029.pdf

Allen, V. (2020). 'What Does Facebook's New Oversight Board Mean for Conservative Posts?'. *The Daily Signal*, 14 May 2020. https://www.dailysignal.com/2020/05/14/what-does-facebooks-new-oversight-board-mean-for-conservative-posts/

Allyn, B. (2020). 'Twitter Flags President Trump's Tweet For The 1st Time'. *NPR*, 26 May 2020. https://www.npr.org/2020/05/26/862838470/twitter-flags-president-trumps-tweet-for-the-first-time

Alves, L. (2018). 'Brazil Preparing to Fight Fake News During October's Elections', *The Rio Times*, 29 June 2018. https://riotimesonline.com/brazil-news/rio-politics/brazil-preparing-to-fight-fake-news-during-octobers-elections/

ANCIR. (n.d). 'Manufacturing Divides: The Gupta-linked Radical Economic Transformation (RET) media network'. https://s3-eu-west-1.amazonaws.com/s3.sourceafrica.net/documents/118115/Manufacturing-Divides.pdf

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P. & Mislove, A. (2018). 'Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations'. *Network and Distributed System Security Symposium*, February 2018. https://hal.archives-ouvertes.fr/hal-01955309/

Annenberg Public Policy Center of the University of Pennsylvania. (2020). 'Freedom and Accountability: A Transatlantic Framework for Moderating Speech Online'. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf

Apelblat, M. (2020). 'UN marks World Press Freedom Day amid attacks against journalists'. *The Brussels Times*, 05 May 2020. https://www.brusselstimes.com/all-news/world-all-news/109683/un-marks-world-press-freedom-day-amid-attacks-against-journalists/

Argentina Political Party Financing Law. (2019). 'Ley De Financiamiento De Los Partidos Políticos Ley 27504'. https://www.boletinoficial.gob.ar/detalleAviso/primera/208603/20190531

Argentinian Government. (2018a). 'Argentina Proyecto de ley Creación de Comisión de Verificación de Noticias Falsas – 2018'. https://www.hcdn.gob.ar/proyectos/textoCompleto.jsp?exp=5228-D-2018&tipo=LEY

Aro, J. (2016). 'The cyberspace war: propaganda and trolling as warfare tools'. *European view* 15.1, 1 June 2016: 121-132. https://journals.sagepub.com/doi/full/10.1007/s12290-016-0395-5

Argentinian Government. (2018b). 'Bill to create a Commission for the Verification of Fake News' (unofficial translation). https://observatoriolegislativocele.com/argentina-proyecto-de-ley-creacion-de-comision-de-verificacion-de-noticias-falsas-2018/

Arthur, R. (2019). 'We Analyzed More Than 1 Million Comments on 4chan. Hate Speech There Has Spiked by 40% Since 2015'. *Vice*, 10 July 2019. https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015

Article 19. (2018a). 'YouTube Community Guidelines: Legal Analysis', September 2018. https://www.article19.org/wp-content/uploads/2018/09/YouTube-Community-Guidelines-August-2018.pdf

Article 19. (2018b). 'Facebook Community Guidelines: Legal Analysis', June 2018. https://www.article19.org/wp-content/uploads/2018/07/Facebook-Community-Standards-August-2018-1-1.pdf

Article 19. (2020a). 'Ensuring the Public's Right to Know in the COVID-19 Pandemic'. May 2020. https://www.article19.org/wp-content/uploads/2020/05/Ensuring-the-Publics-Right-to-Know-in-the-Covid-19-Pandemic_Final.pdf

Article 19. (2020b). 'We share the concern of many at reports today of the intention of the US President to sign an executive order that will seek to curtail the free speech protections....' *Twitter*, 28 May 2020. https://twitter.com/article19org/status/1266017734925656065?s=20

ASEAN. (2020). 'Statement of the Special ASEAN-China Foreign Ministers' Meeting on the Coronavirus Disease 2019 (COVID-19)'. 20 February 2020. https://asean.org/storage/2020/02/ASEAN-China-SFMM-Statement-on-COVID-19-20-Feb-2020-Final.pdf

ASEAN AMRI. (2018). 'The ASEAN Ministers Responsible for Information (AMRI) Framework and Joint Declaration to Minimise the Harmful Effects of Fake News'. 10 May 2018. https://asean.org/storage/2012/05/Annex-5-Framework-Declr-Fake-News.pdf

Aspray, W. & Cortada, J.W. (2019). 'From Urban Legends to Political Fact-Checking'. Springer.

Associated Press. (2018). '3 Myanmar journalists in court over story gov't calls false'. 17 October 2018. https://apnews.com/e7d951b784ac48208000d2609435dbab?utm_source=Pew+Research+Center&utm_campaign=c32d4eb996-EMAIL_CAMPAIGN_2018_10_18_01_28&utm_medium=email&utm_term=0_3e953b9b70-c32d4eb996-400451153

Associated Press. (2019). 'Bahrain charges lawyer of sharing "fake news" for his Tweets'. *The Washington Post*, 15 May 2019. https://www.washingtonpost.com/world/middle_east/bahrain-charges-lawyer-of-sharing-fake-news-for-his-tweets/2019/05/15/524e4a04-7727-11e9-a7bf-c8a43b84ee31_story.html

Avaaz. (2019). 'Far Right Networks of Deception'. 22 May 2019. https://avaazimages.avaaz.org/Avaaz%20Report%20Network%20Deception%2020190522.pdf?slideshow

Avaaz. (2020). 'Facebook's Algorithm: A Major Threat to Public Health'. 19 August 2020. https://secure.avaaz.org/campaign/en/facebook_threat_health/

Axelrod, T. (2020). 'Facebook donating $2M to local newsrooms, fact-checkers covering coronavirus'. *The Hill*, 17 March 2020. https://thehill.com/policy/technology/488065-facebook-donating-1-million-to-local-newsrooms-covering-coronavirus-and

Ayyub, R. (2018). "In India, journalists face slut-shaming and rape threats." *New York Times*, 22 May 2018. https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html

Babakar, M. & Moy, W. (2016) 'The State of Automated Factchecking'. *Full Fact report*, 17 August 2016. https://fullfact.org/blog/2016/aug/automated-factchecking/

Babu, A., Lui, A., & Zhang, J. (2017). 'New Updates to Reduce Clickbait Headlines', *Facebook Newsroom*, 17 May 2017. https://newsroom.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

Ball, J. (2018). 'Post-truth: How bullshit conquered the world'. *Biteback Publishing*, 2017.

Bälz, K. & Mujally, H. (2019). 'Egypt: The New Egyptian Anti-Cybercrime Law Regulates Legal Responsibility For Web Pages And Their Content'. Mondaq, 1 July 2019. http://www.mondaq.com/x/—820028/Security/The+New+Egyptian+AntiCybercrime+Law+Regulates+Legal+Responsibility+for+Web+Pages+and+Their+Content

Bangkok Post. (2019). 'Anti-fake news centre hails first arrest'. 14 November 2019. https://www.bangkFokpost.com/thailand/general/1793649/anti-fake-news-centre-hails-first-arrest

Bangladesh Digital Security Act. (2018). Act No 46 of the Year 2018. 08 October 2018. https://www.cirt.gov.bd/wp-content/uploads/2018/12/Digital-Security-Act-2018-English-version.pdf

Barnathan, J. (2020). 'A Global Crisis Like COVID-19 Calls for a Global Response. Here's Ours.' *ICFJ*, 20 March 2020. https://www.icfj.org/news/global-crisis-covid-19-calls-global-response-heres-ours

Barnett, S. (2016). 'How our mainstream media failed democracy'. In EU Referendum Analysis 2016: Media, Voters and the Campaign. Jackson, D., Thorsen, E., Wring, D., Loughborough University Center for the Study of Journalism, Culture and Community. http://www.referendumanalysis.eu/

Baron Cohen, S. (2019). 'Never Is Now 2019. ADL International Leadership Award Presented to Sacha Baron Cohen'. Recording of his speech, *YouTube*, 21 November 2019. https://www.youtube.com/watch?v=ymaWq5yZIYM&feature=youtu.be

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2014). 'Anti-social media'. *Technical report, Demos*, February 2014. https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf

BBC. (2018a). 'Bahrain activist jailed for five years over Twitter comments'. 21 February 2018. https://www.bbc.co.uk/news/world-middle-east-43140519

BBC. (2018b). 'Egypt sentences activist for 'spreading fake news''. 29 September 2018. https://www.bbc.co.uk/news/world-middle-east-45691770

BBC. (2018c). 'Beyond Fake News. BBC launches huge new international anti-disinformation initiative'. 09 November 2018. https://www.bbc.co.uk/mediacentre/latestnews/2018/beyond-fake-news

BBC. (2019a). 'Ivorian MP Alain Lobognan jailed for 'fake news' tweet'. 30 January, 2019. https://www.bbc.co.uk/news/world-africa-47057509

BBC. (2019b). 'Russia internet: Law introducing new controls comes into force'. 01 November 2019. https://www.bbc.co.uk/news/world-europe-50259597

BBC. (2020a). 'BBC Young Reporter and British Council team up to fight 'fake news''. 16 January 2020. https://www.bbc.co.uk/mediacentre/latestnews/2020/young-reporter-fake-news

BBC. (2020b). 'Coronavirus: World leaders' posts deleted over fake news'. 31 March 2020. https://www.bbc.co.uk/news/technology-52106321

BBC. (2020c). 'Twitter hides Trump tweet for 'glorifying violence'. 29 May 2020. https://www.bbc.co.uk/news/technology-52846679

BBC. (2020d). 'Facebook and Twitter restrict Trump accounts over 'harmful' virus claim'. 06 August 2020. https://www.bbc.co.uk/news/election-us-2020-53673797

Beavers, O. (2019). ' Experts are studying mannerisms of 2020 candidates to help offset threat of 'deepfake' videos'. *The Hill*, 09 May 2019. https://thehill.com/policy/cybersecurity/443018-experts-are-studying-mannerisms-of-2020-candidates-to-help-offset-threat

Beckett, C. (2016). 'Deliberation, distortion and dystopia: the news media and the referendum'. *In EU Referendum Analysis 2016: Media, Voters and the Campaign. Jackson, D, Thorsen, E, Wring, D. Loughborough University Center for the Study of Journalism, Culture and Community*. http://www.referendumanalysis.eu/

Bell, E. & Owen, T. (2017). 'The Platform Press: How Silicon Valley Reengineered Journalism'. *Tow Center for Digital Journalism*, 29 March 2017. https://www.cjr.org/tow_center_reports/platform-press-how-silicon-valley-reengineered-journalism.php

Bengani, P. (2020). 'As election looms, a network of mysterious 'pink slime' local news outlets nearly triples in size'. *CJR*, 4 August 2020. https://www.cjr.org/analysis/as-election-looms-a-network-of-mysterious-pink-slime-local-news-outlets-nearly-triples-in-size.php

Benin Digital Code, (2017). Assemblee Nationale. Loi n° 2017-20 portant code du numérique en République du Bénin. 13 June 2017. https://www.afapdp.org/wp-content/uploads/2018/06/Benin-Loi-2017-20-Portant-code-du-numerique-en-Republique-du-Benin.pdf

Benkler, Y., Faris, B., Roberts, H. (2018). 'Network Propaganda. Manipulation, Disinformation, and Radicalization in American Politics'. *Oxford University Press*.

Benton, J. (2020). 'Is this video "missing context," "transformed," or "edited"? This effort wants to standardize how we categorize visual misinformation'. *NiemanLab*, 16 January 2020. https://www.niemanlab.org/2020/01/is-this-video-missing-context-transformed-or-edited-this-effort-wants-to-standardize-how-we-categorize-visual-misinformation/

Beo Da Costa, A. (2019). 'Indonesia lifts social media curbs targeting hoaxes during unrest'. *Reuters*, 25 May 2019. https://www.reuters.com/article/us-indonesia-election-social-media/indonesia-lifts-social-media-curbs-targeting-hoaxes-during-unrest-idUSKCN1SV06J

Berger, G. (2019). 'Whither MIL: Thoughts for the Road Ahead'. Chapter in 'Understanding Media and Information Literacy (MIL) in the Digital Age. A Question of Democracy'. (ed). Ulla Carlsson, Goteborg: Nordicom. https://jmg.gu.se/digitalAssets/1742/1742676_understanding-media-pdf-original.pdf

Bhattacharjee, S. & Dotto, C. (n.d.) 'First Draft case study: Understanding the impact of polio vaccine disinformation in Pakistan'. *First Draft*. https://firstdraftnews.org/long-form-article/first-draft-case-study-understanding-the-impact-of-polio-vaccine-disinformation-in-pakistan/

Bich Ngoc, N. (2019). 'Vietnam's New Cybersecurity Law 2018'. *Vietnam Business Law*. 30 July 2018. https://vietnam-business-law.info/blog/2018/7/30/vietnams-new-cybersecurity-law

Bickert, M. (2018). 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', *Facebook Newsroom*, 24 April 2018. https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/

Bickert, M. (2019). 'Updating the Values That Inform Our Community Standards', *Facebook Newsroom*, 12 September 2019. https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/

Bigot, L. (2019). Fact-Checking vs. fake news. Vérifier pour mieux informer. *INA*, 18 October 2019. https://presse.ina.fr/fact-checking-vs-fake-news/

Billing, L. (2020). 'Duterte's troll armies drown out Covid-19 dissent in the Philippines'. *Coda*, 21 July 2020. https://www.codastory.com/disinformation/philippines-troll-armies/

Bilton, R. (2016). 'Electionland, a joint project between ProPublica and six other orgs, will cover Election Day voting issues'. *NiemanLab*, 08 September 2016. https://www.niemanlab.org/2016/09/electionland-a-joint-project-between-propublica-and-six-other-orgs-will-create-a-virtual-newsroom-to-cover-election-day-voting-issues/

Binder, M. (2019). 'Facebook ad scam tricks users with images and video of Kickstarter products'. *Mashable UK*, 15 November 2019. https://mashable.com/article/facebook-scam-crowdfunding-ads.amp/?europe=true

Birks, J. (2019). '*Fact-Checking Journalism and Political Argumentation. A British Perspective*'. Palgrave / MacMillan

Board, J. (2019). 'Inside Indonesia's 'fake news' war room, fighting political hoaxes in election season'. *CAN*, 13 April 2019. https://www.channelnewsasia.com/news/asia/indonesia-election-fake-news-war-room-fighting-political-hoaxes-11439398

Boothroyd-Rojas, R. (2017). 'Venezuela's Constituent Assembly Passes Anti-Hate Crime Law'. *Venezuelanalysis*, 09 November 2017. https://venezuelanalysis.com/news/13492

Bosworth, A. (2020). 'Lord of the Rings, 2020 and Stuffed Oreos: Read the Andrew Bosworth Memo'. *New York Times*, 7 January 2020. https://www.nytimes.com/2020/01/07/technology/facebook-andrew-bosworth-memo.html

Bossetta, M. (2018). 'The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election'. *Journalism & Mass Communication Quarterly*, 28 March 2018. https://doi.org/10.1177/1077699018763307

Bozdag, E., & Van den Hoven, M. J. (2015). 'Breaking the filter bubble: Democracy and design'. *Ethics and Information Technology. 17.4, 249-265*, 18 December 2015. https://repository.tudelft.nl/islandora/object/uuid%3Ae24751ba-b94b-4856-b4a9-2f5f4f25ff14

Bradshaw, S. & Howard, P.N. (2018). 'Challenging Truth and Trust: a Global Inventory of Organized Social Media Manipulation'. *Working Paper 2018.1. Oxford, UK: Project on Computational Propaganda, 26 pp*, 20 July 2018. https://comprop.oii.ox.ac.uk/research/cybertroops2018/

Bradshaw, S. & Howard, P.N. (2019). 'The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation'. *Oxford, UK: Project on Computational Propaganda, 26 September 2019*. https://comprop.oii.ox.ac.uk/research/cybertroops2019/

Brazil Superior Electoral Court. (2018). 'Fake News: TSE lança página para esclarecer eleitores', Press Release, 11 October 2018. http://www.tse.jus.br/imprensa/noticias-tse/2018/Outubro/fake-news-tse-lanca-pagina-para-esclarecer-eleitores-sobre-a-verdade

Brennen, J. S., Simon, F., Howard, P. N. & Nielsen, R. K. (2020). 'Types, sources, and claims of COVID-19 misinformation'. *Reuters*, 07 April 2020. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Broadband Commission. (2013). Technology, broadband and education: Advancing the education for all agenda. *UNESCO Publication*. https://unesdoc.unesco.org/ark:/48223/pf0000219687.locale=en

Brooking, E.T., Kann, A. & Rizzuto, M. (2020). 'Dichotomies of Disinformation'. *DFRLab*. https://github.com/DFRLab/Dichotomies-of-Disinformation/blob/master/README.md

Buchanan, K. (2019). 'Initiatives to Counter Fake News: Malaysia'. *Library of Congress*, April 2019. https://www.loc.gov/law/help/fake-news/malaysia.php

Buchanan, T. & Benson, V. (2019). 'Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"?'. *Social Media + Society*, 17 December 2019. https://doi.org/10.1177/2056305119888654

Buckmaster, L. & Wils, T. (2019). 'Responding to Fake News. *Australia Parliamentary Library Briefing*'. https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/BriefingBook46p/FakeNews

Bukhari, P. (2019). 'Srinagar, India | AFP | Monday 12/23/2019 – 15:28 UTC+8 | 663 words'. *The Independent News Singapore*, 23 December 2019. http://theindependent.sg/srinagar-india-afp-monday-12-23-2019-1528-utc8-663-words/

Buning, M.D.C. et al. (2018). 'A multi-dimensional approach to disinformation'. Report of the independent High Level Group on Fake News and Online Disinformation'. *European Commission*, 12 March 2018. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Burgess, M. (2018). 'To fight fake news on WhatsApp, India is turning off the internet'. *Wired*, 18 October 2018. https://www.wired.co.uk/article/whatsapp-web-internet-shutdown-india-turn-off

Burgos, P. (2019). 'What 100,000 WhatsApp Messages Reveal about Misinformation in Brazil', *First Draft*, 27 June 2019. https://firstdraftnews.org/latest/what-100000-whatsapp-messages-reveal-about-misinformation-in-brazil/

Burki, T. (2019). 'Vaccine misinformation and social media'. *The Lancet*, 01 October 2019. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30136-0/fulltext

Burkina Faso Constitutional Council. (2019). 'Décision n° 2019- 013/CC sur le contrôle de constitutionnalité de la loi n° 044-2019/AN du 21 juin 2019 portant modification de la loi n° 025-2018/AN du 31 mai 2018 portant Code pénal par autosaisine'. https://www.conseil-constitutionnel.gov.bf/fileadmin/user_upload/decision_13__code_penal.pdf

Cadwalladr, C. (2017a). Revealed: Tory 'dark' ads targeted voters' Facebook feeds in Welsh marginal seat. *The Observer*. 27 May 2017. https://www.theguardian.com/politics/2017/may/27/conservativesfacebook-dark-ads-data-protection-election

Cadwalladr, C. (2017b). 'The great British Brexit robbery: how our democracy was hijacked'. *The Guardian*, 07 May 2017. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

Cadwalladr, C. (2018). ''Plucky little panel' that found the truth about fake news Facebook and Brexit'. *The Guardian*, 28 Jul 2018. https://www.theguardian.com/politics/2018/jul/28/dcms-committee-reportfinds-truth-fake-news-facebook-brexit

Cameroon Cyber Security and Cyber Criminality Law. (2010). 'Loi n°2010/012 du 21 Decembre 2010 relative a la Cybersecurite et la Cybercriminalite au Cameroun'. https://www.unodc.org/res/cld/document/cmr/2010/loi_sur_la_cybersecurite_et_la_cybercriminalite_html/Loi_2010-012_cybersecurite_cybercriminalite.pdf

Cameroon Penal Code (1967 Revised). https://www.wipo.int/edocs/lexdocs/laws/en/cm/cm014en.pdf

Canada House of Commons Standing Committee on Access to Information, Privacy and Ethics. (2018). 'Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly'. *Ottowa*: *House of Commons*. https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-17

Canada House of Commons Standing Committee on Access to Information, Privacy and Ethics. (2019). *International Grand Committee on Big Data, Privacy and Democracy. Report and Government Response*. https://www.ourcommons.ca/Committees/en/ETHI/StudyActivity?studyActivityId=10554743

Canadian Government. (2018). 'Elections Modernization Act'. https://laws-lois.justice.gc.ca/eng/annualstatutes/2018_31/page-1.html

Canadian Government. (2019a). *Response to Canadian Parliamentary Standing Committee on Access to Information, Privacy and Ethics Report. Democracy under Threat*: *Risks and Solutions in the Era of Disinformation and Data Monopoly*. https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-17/response-8512-421-502

Canadian Government. (2019b). *Online Disinformation*. https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html

Canadian Government. (2019c). 'Backgrounder – Helping Citizens Critically Assess and Become Resilient Against Harmful Online Disinformation'. *Canadian Heritage*, 21 August 2019. https://www.canada.ca/en/canadian-heritage/news/2019/07/backgrounder--helping-citizens-critically-assess-and-become-resilient-against-harmful-online-disinformation.html

Canadian Heritage. (2020). 'COVID-19: The Government of Canada is taking action to support the publishing and news sectors'. https://www.canada.ca/en/canadian-heritage/news/2020/03/covid-19-the-government-of-canada-is-taking-action-to-support-the-publishing-and-news-sectors.html

Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B. & Zeitoff, T. (2020). 'The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil'. Science Advances, January 2020. DOI: 10.1126/sciadv.aaw7449

Carmi, E., Yates, S.J., Lockley, E. & Pawluczuk, A. (2020). 'Data citizenship: rethinking data literacy in the age of disinformation, misinformation, and malinformation'. *Internet Policy Review*, 28 May 2020. https://policyreview.info/articles/analysis/data-citizenship-rethinking-data-literacy-age-disinformation-misinformation-and

Carmichael, F. & Gragnani, J. (2019). 'YouTube advertises big brands alongside fake cancer cure videos'. *BBC*, 13 September 2019. https://www.bbc.co.uk/news/blogs-trending-49483681

Carmichael, F. & Hussain, A. (2019). 'Pro-Indian 'fake websites targeted decision makers in Europe'. *BBC*, 16 December 2019. https://www.bbc.co.uk/news/world-asia-india-50749764

Cassini, S. (2019). 'Christophe Castaner et la Pitié-Salpêtrière : premier désaveu pour la loi sur les « infox »'. *Le Monde*, 22 May 2019. https://www.lemonde.fr/politique/article/2019/05/22/premier-desaveu-pour-la-loi-sur-les-infox_5465717_823448.html

Centre for Data Ethics and Innovation. (2019). 'Deepfakes and Audio-visual Disinformation'. September 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831179/Snapshot_Paper_-_Deepfakes_and_Audiovisual_Disinformation.pdf

Center for Human Rights, United Nations. (1994). 'Handbook on the Legal, Technical, and Human Rights Aspects of Elections'. https://eos.cartercenter.org/uploads/document_file/path/8/training2enTCCoptimized.pdf

Chakraborty, A., Sarkar, R., Mrigen, A., & Ganguly, N. (2017). 'Tabloids in the Era of Social Media? Understanding the Production and Consumption of Clickbaits in Twitter'. SSRN, 10 July 2017. http://dx.doi.org/10.2139/ssrn.3034591

Chaturvedi, S. (2016). 'I am a troll: Inside the secret world of the BJP's digital army'. *Juggernaut Books*, 2016.

Chaturvedi, S. (2020). 'Govt launches chatbot on WhatsApp to create awareness about coronavirus, curb misinformation'. *The Economic Times*, 22 March 2020. https://economictimes.indiatimes.com/tech/internet/govt-launches-chatbot-on-whatsapp-to-create-awareness-about-coronavirus-curb-misinformation/articleshow/74750648.cms

Chavoshi, N., Hamooni, H., & Mueen, A. (2017). 'Temporal Patterns in Bot Activities'. *ACM*, April, 2017. http://doi.org/10.1145/3041021.3051114

Checknews (2019). '*Combien a rapporté à Libé son partenariat de factchecking avec Facebook en 2018*?'. 30 January 2019. https://www.liberation.fr/checknews/2019/01/30/combien-a-rapporte-a-libe-son-partenariat-de-factchecking-avec-facebook-en-2018_1706160

Chile Senate. (2019). "Fake news": autoridades nacionales y regionales podrían tener nueva causal para la cesación de cargos. Press Release, 7 February 2019. https://www.senado.cl/fake-news-autoridades-nacionales-y-regionales-podrian-tener-nueva/senado/2019-01-29/122725.html

Chitranukroh, A. (2017). 'The new Computer Crimes Act and concerns over online freedom'. *Bangkok Post*, 20 January 2017. https://www.bangkokpost.com/business/1183561/the-new-computer-crimes-act-and-

Christopher, N. (2020). 'We've Just Seen the First Use of Deepfakes in an Indian Election Campaign'. *Vice*, 18 February 2020. https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Chung Seck, Y. & Son Dang, T. (2019). 'Vietnam National Assembly Passes the Law on Cybersecurity'. *Global Compliance News*, 02 July 2018. https://globalcompliancenews.com/vietnam-law-cybersecurity-20180702/

Clegg, N. (2019). 'Facebook, Elections and Political Speech'. *Facebook*, 24 September 2019. https://about.fb.com/news/2019/09/elections-and-political-speech/

Clegg, N. (2020). 'Welcoming the Oversight Board'. *Facebook*, 06 May 2020. https://about.fb.com/news/2020/05/welcoming-the-oversight-board/

Cohen, N. (2020). 'How Wikipedia Prevents the Spread of Coronavirus Misinformation'. *Wired*, 15 March 2020. https://www.wired.com/story/how-wikipedia-prevents-spread-coronavirus-misinformation/

Collingridge, D. (1980). 'The social control of technology'. St Martin, New York

Conseil Constitutionnel. (2018). 'Décision n° 2018-773 DC du 20 décembre 2018'. https://www.conseil-constitutionnel.fr/decision/2018/2018773DC.htm

Cook, J., Ecker, U.K.H & Lewandowsky, S. (2014). 'Misinformation and How to Correct It.' *ResearchGate*, July 2014. https://www.researchgate.net/publication/265643150_Misinformation_and_How_to_Correct_It

Cook, J., van der Linden, S., Lewandowsky, S. & Ecker, U. (2020). 'Coronavirus, 'Plandemic' and the seven traits of conspiratorial thinking'. *The Conversation*, 15 May 2020. https://theconversation.com/coronavirus-plandemic-and-the-seven-traits-of-conspiratorial-thinking-138483

Corbu, N., Oprea, D-A., Negrea-Busuioc, E. & Radu, L. (2020). ''They can't fool me, but they can fool the others!' Third person effect and fake news detection'. *European Journal of Communication*, 17 February 2020. https://journals.sagepub.com/doi/full/10.1177/0267323120903686?journalCode=ejca

Corcoran, C. Crowley, B.J., Davis, R. (2019) 'Disinformation Threat Watch. The Disinformation Landscape in East Asia and Implications for US Policy'. *Harvard Kennedy School, Student Report*, May 2019. https://www.belfercenter.org/sites/default/files/2019-06/PAE/DisinfoWatch%20-%202.pdf

Cordey, S. (2019). 'Cyber Influence Operations: An Overview and Comparative Analysis'. *Zurich*: *Center for Security Studies* (*CSS*). https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/Cyber-Reports-2019-10-CyberInfluence.pdf

Côte d'Ivoire Penal Code 1981, revised. Article 97, 2017. http://www.caidp.ci/uploads/01981c9a7d883c4321811e8725ca4c2c.pdf

Côte d'Ivoire Penal Code 1981, revised. Article 173. http://www.gouv.ci/doc/accords/1512502410CODE-PENAL.pdf

CPJ. (2019a). 'Cameroonian journalist detained on criminal defamation and false news charges'. 20 June 2019. https://cpj.org/2019/06/critical-cameroonian-journalist-detained-on-crimin.php

CPJ. (2019b). 'Journalist Ignace Sossou convicted of false news in Benin'. 23 August 2019. https://cpj.org/2019/08/journalist-ignace-sossou-convicted-of-false-news-i.php

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. (2016). 'DNA-inspired online behavioral modeling and its application to spambot detection'. *IEEE Intelligent Systems*, 31(5), 30 January 2016. https://www.researchgate.net/publication/298902745_DNA-Inspired_Online_Behavioral_Modeling_and_Its_Application_to_Spambot_Detection

Damiano Ricci, A. (2018). 'French opposition parties are taking Macron's anti-misinformation law to court'. *Poynter*, 04 December 2018. https://www.poynter.org/fact-checking/2018/french-opposition-parties-are-taking-macrons-anti-misinformation-law-to-court/

Danish Ministry of Foreign Affairs. (2018). 'Strengthened safeguards against foreign influence on Danish elections and democracy'. Press Release, 07 September 2018. https://um.dk/en/news/newsdisplaypage/?newsid=1df5adbb-d1df-402b-b9ac-57fd4485ffa4

Dara, V. (2019). 'Government to launch TV show against fake news'. *The Phnom Penh Post*, 24 January 2019. https://www.phnompenhpost.com/national/government-launch-tv-show-against-fake-news

Darmanin, J. (2019). 'FCEU Newsletter #7 - Good news and bad news after the election week-end'. *Fact Check EU*, 27 May 2019. https://factcheckeu.info/en/article/fceu-newsletter-7-good-news-and-bad-news-after-election-week-end

Daudin, G. (2019). '52% des viols commis à Paris en 2014 l'ont été par des étrangers. L'étude ne dit pas cela'. *AFP Factuel*, 25 November 2019. https://factuel.afp.com/52-des-viols-commis-paris-en-2014-lont-ete-par-des-etrangers-letude-ne-dit-pas-cela

Davis, A. (2019). Political Communication. A New Introduction for Crisis Times. Polity.

Dawn. (2018). 'Govt launches 'Fake News Buster' account to expose false reports'. 01 October 2018. https://www.dawn.com/news/1436167

Dean, M. (2017). 'Snopes and the Search for Facts in a Post-Fact World'. *Wired*, 20 September 2017. https://www.wired.com/story/snopes-and-the-search-for-facts-in-a-post-fact-world/

Decker, B. (2019). 'Adversarial Narratives: A New Model for Disinformation'. *GDI*, August 2019. https://disinformationindex.org/wp-content/uploads/2019/08/GDI_Adverserial-Narratives_Report_V6.pdf

DCMS HC 363 (2018). 'Oral evidence : Fake News – 8 February 2018 (George Washington University, Washington DC), HC 363'. 08 February 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/fake-news/oral/78195.html

De Croo, A. (2018). 'Fact checking fonds in de steigers in strijd tegen fake news'. Press Release, 08 October 2018. https://alexanderdecroo.be/fact-checking-fonds-steigers-strijd-fake-news/

Derakhshan, D. (2019). 'Disinfo Wars. A taxonomy of information warfare'. *Medium*, 9 May 2019. https://medium.com/@h0d3r/disinfo-wars-7f1cf2685e13

Dewey, C. (2016). 'Facebook Fake-News Writer: 'I Think Donald Trump is in the White House because of Me.'' *The Washington Post*, November 17, 2016. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/

DiResta, R. (2018). 'Free Speech Is Not the Same As Free Reach'. *Wired*, 30 August 2018. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/

Dobber, T., Ó Fathaigh, R. & Zuiderveen Borgesius, F. (2019). 'The regulation of online political micro-targeting in Europe'. *Internet Policy Review*, 31 December 2019. https://policyreview.info/articles/analysis/regulation-online-political-micro-targeting-europe

Dorroh, J. (2020). Beyond fact-checking: fighting the onslaught of COVID-19 disinformation. Ijnet international journalists' network, 2 July 2020. https://ijnet.org/en/story/beyond-fact-checking-fighting-onslaught-covid-19-disinformation

Dorsey, J. (2019). 'We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought. Why? A few reasons...' *Twitter*, 30 October 2019. https://twitter.com/jack/status/1189634360472829952?s=20

Douek, E. (2020). 'COVID-19 and Social Media Content Moderation'. *Lawfare*, 25 March 2020, https://www.lawfareblog.com/covid-19-and-social-media-content-moderation

Doyle, A. (2016). 'Facebook Says Will Learn From Mistake Over Vietnam Photo'. *Hindustan Times*, 13 September 2016. http://www.hindustantimes.com/world-news/facebook-says-will-learn-from-mistake-over-vietnam-photo/story-kwmb3iX6lKgmwaIGZeKlyN.html

Drissa, D. (2019). 'Offensive générale contre les fausses informations sur les réseaux sociaux'. *7info*, 12 July 2019. https://www.7info.ci/offensive-generale-contre-les-fausses-informations-sur-les-reseaux-sociaux/

Drobnic Holan, A. (2018). 'Politifact: The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking'. *Politifact*, 12 February 2018. https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

Duarte, N. & Llansó, E. (2017). 'Mixed Messages: The Limits of Automated Social Media Content Analysis'. *Center for Democracy and Technology*, 28 November 2017. https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/

Dubois, E., & Blank, G. (2018). 'The echo chamber is overstated: the moderating effect of political interest and diverse media'. Information, Communication & Society, 29 January 2018. 21(5), 729-745. https://doi.org/10.1080/1369118X.2018.1428656

Dufour, N. & Gully, A. (2019). 'Contributing Data to Deepfake Detection Research'. *Google Blog*, 24 September 2019. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

Durach, F. (2020). 'Who's afraid of fake news? New evidence from Romania'. *YouCheck*, 24 February 2020. http://project-youcheck.com/whos-afraid-of-fake-news-new-evidence-from-romania/

Dutch Government (2019a). Fake news campaign starts today. Press Release, 11 March 2019. https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws/nieuws/2019/03/11/campagne-nepnieuws-vandaag-van-start

Dutch Government (2019b). Kabinet zet in op transparantie in strategie tegen desinformatie. Press Release, 18 October 2019). https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws/nieuws/2019/10/18/kabinet-zet-in-op-transparantie-in-strategie-tegen-desinformatie

Dutch Government (2019c). Actielijnen tegengaan desinformatie. https://www.rijksoverheid.nl/onderwerpen/desinformatie-nepnieuws/documenten/kamerstukken/2019/10/18/actielijnen-tegengaan-desinformatie

Dwoskin, E., Whalen, J. & Cabato, R. (2019). 'Content moderators at YouTube Facebook and Twitter see the worst of the web – and suffer silently. *The Washington Post*, 25 July 2019. https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/

EC Communication on Tackling Illegal Content Online. (COM (2017) 555 final) https://webcache.googleusercontent.com/search?q=cache:ZJcX6Dn0020J:https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-555-F1-EN-MAIN-PART-1.PDF+&cd=1&hl=en&ct=clnk&gl=uk&client=firefox-b-e

ECI. (2019). 'Report of the Committee on Section 126 of the Representation of the People Act, 1951 Submitted to The Commission'. 10 January 2019. https://pib.gov.in/newsite/PrintRelease.aspx?relid=187412

Ecker, U. K. H., Lewandowsky, S, and Tang, D.T.W. (2010). "Explicit warnings reduce but do not eliminate the continued influence of misinformation". *Memory & Cognition*. Vol.38, No. 8, pp. 1087-1100 https://doi.org/10.3758/MC.38.8.1087

Ecker, U. K. H., O'Reilly, Z., Reid, J. S. & Chang, E. P. (2019). 'The effectiveness of short-format refutational fact-checks'. *British Journal of Psychology*, 02 March 2019. https://onlinelibrary.wiley.com/doi/full/10.1111/bjop.12383

Eco, U. (2014). 'From the Tree to the Labyrinth'. *Harvard University Press*.

Effron, D. A. & Raj, M. (2019). 'Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share'. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

EFJ (2018). 'Belarus: more media censorship and control with new amendments of Media Law'. 24 June 2018. https://europeanjournalists.org/blog/2018/06/24/belarus-more-media-censorship-and-control-with-new-amendments-of-the-media-law/

Eisenstat, Y. (2019). 'I worked on political ads at Facebook. They profit by manipulating us.' The Washington Post, 04 November 2019. https://www.washingtonpost.com/outlook/2019/11/04/i-worked-political-ads-facebook-they-profit-by-manipulating-us/

El Khoury, R. (2020). 'WhatsApp has rolled out the restriction of single chat forwards for viral messages'. *Android Police*, 13 April 2020. https://www.androidpolice.com/2020/04/13/whatsapp-will-limit-viral-message-forwards-to-one-chat-at-a-time-to-combat-coronavirus-misinformation/

Ellis, S. (1989). 'Tuning in to pavement radio.' *African Affairs* 88.352, July 1989 pp321-330 https://pdfs.semanticscholar.org/d07a/60d84e60248e9523c213c7103cd27ce84f8e.pdf

Ellis-Petersen, H. (2019). 'Social media shut down in Sri Lanka in bid to stem misinformation'. *The Guardian, 21* April 2019. https://www.theguardian.com/world/2019/apr/21/social-media-shut-down-in-sri-lanka-in-bid-to-stem-misinformation

Emarketer. (2020). 'US Political Ad Spending to Hit Record High'. 12 February 2020. https://www.emarketer.com/newsroom/index.php/us-political-ad-spending-to-hit-record-high/

Embury-Dennis, T. (2020). 'Coronavirus: New York sees spike in disinfectant exposure cases following Trump's dangerous treatment musings'. *Independent*, 25 April 2020. https://www.independent.co.uk/news/world/americas/us-politics/coronavirus-trump-treatment-disinfectant-bleach-new-york-a9483786.html

Endeshaw, D. (2020). 'Ethiopia passes law imposing jail terms for internet posts that stir unrest'. *Reuters*, 13 February 2020. https://www.reuters.com/article/us-ethiopia-politics/ethiopia-passes-law-imposing-jail-terms-for-internet-posts-that-stir-unrest-idUSKBN2071PA

Election Monitoring. (2019). 'UK General Election 2019; Digital disruption by the political parties, and the need for new rules'. December, 2019. https://www.isdglobal.org/wp-content/uploads/2019/12/UK-GE-2019-Digital-Disruption-report.pdf

Eljechtimi, A. (2020). 'Morocco makes a dozen arrests over coronavirus fake news'. *Reuters*, 19 March 2020. https://uk.reuters.com/article/uk-health-coronavirus-morocco/morocco-makes-dozen-arrests-over-coronavirus-fake-news-idUKKBN2162EA

Epstein R., Robertson, R. E. (2015). 'The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections'. In *Proceedings of the National Academy of Sciences* (*PNAS*). *112*(*33*):*E4512–E4521*, 18 August 2015. https://www.pnas.org/content/112/33/E4512

Estarque, M. (2020). 'Brazilian projects for media literacy and combating false news find allies outside journalism'. *Knight Center*, 29 January 2020. https://knightcenter.utexas.edu/blog/00-21557-brazilian-projects-media-literacy-and-combating-false-news-find-allies-outside-journal

EU Disinfo Lab (2019a). 'Uncovered: 265 coordinated fake local media outlets serving Indian interests'. 26 November 2019. https://www.disinfo.eu/publications/uncovered-265-coordinated-fake-local-media-outlets-serving-indian-interests

EU Disinfo Lab (2019b). 'An investigation into a pro-Indian influence network'. https://www.disinfo.eu/wp-content/uploads/2019/12/20191213_InfluencingPolicymakers-with-Fake-media-outlets.pdf

EU Disinfo Lab (2019c). 'How you thought you support the animals and you ended up funding white supremacists'. 11 September 2019. https://www.disinfo.eu/publications/suavelos-white-supremacists-funded-through-facebook

EU Disinfo Lab. (2020). 'From health disinformation to copy-pasting Sputnik and RT articles – how an Africa-based network built fake media outlets and clickbait websites for profit'. 8 March 2020. https://www.disinfo.eu/publications/from-health-disinformation-to-copy-pasting-articles-from-sputnik-and-rt-how-an-africa-based-network-built-fake-media-outlets-and-clickbait-websites-for-profit

EU EEAS (2018). 'Questions and Answers about the East Stratcom Task Force'. https://eeas.europa.eu/headquarters/headquarters-homepage/2116/-questions-and-answers-about-the-east-stratcom-task-force_en

EU Foreign Affairs Council. (2018). 'Outcome of The Council Meeting'. 16 April 2018. https://www.consilium.europa.eu/media/33743/st07997-en18.pdf

EU HLEG. (2018). EU High Level Expert Group on Fake News and Online Disinformation. '*A Multi-Dimensional Approach to Disinformation*'. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

EurActiv. (2016). 'Latvia shuts down Russian 'propaganda' website Sputnik'. 30 March 2016. https://www.euractiv.com/section/global-europe/news/latvia-shuts-down-russias-propaganda-website-sputnik/

EurActiv. (2019). 'France adopts tough law against online hate speech'. 10 July 2019. https://www.euractiv.com/section/politics/news/france-adopts-tough-law-against-online-hate-speech/

Eurobarometer 464. (2018). 'Final results of the Eurobarometer on fake news and online disinformation'. https://ec.europa.eu/digital-single-market/en/news/final-results-eurobarometer-fake-news-and-onlinedisinformation

European Commission. (2018a). 'Communication - Tackling online disinformation: a European Approach' (COM(2018) 236 final), 26 April 2018. https://ec.europa.eu/digital-single-market/en/news/communication-tackling-online-disinformation-european-approach

European Commission. (2018b). 'State of the Union 2018: European Commission Proposes Measures for Securing Free and Fair European Elections', Press Release (IP/18/5681), 12 September 2018. http://europa.eu/rapid/press-release_IP-18-5681_en.htm

European Commission. (2018c). 'Code of Practice on Disinformation'. 26 September 2018. https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Commission. (2018d). 'Roadmaps to implement the Code of Practice on disinformation'. 16 October 2018. https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation

European Commission. (2018e) Joint Communication. '*Action Plan against Disinformation*' *JOIN (2018) 36 final*, 05 December 2018. https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation

European Commission. (2019). Policy 'Tackling Online Disinformation'. 13 September 2019. https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation

European Commission and High Representative. (2018). 'Action Plan against Disinformation'. 05 December 2018. https://ec.europa.eu/commission/sites/beta-political/files/eu-communication-disinformation-euco-05122018_en.pdf

European Parliament. (2020). 'Disinformation: how to recognise and tackle Covid-19 myths'. 30 March 2020. https://www.europarl.europa.eu/news/en/headlines/society/20200326STO75917/disinformation-how-to-recognise-and-tackle-covid-19-myths

EUvsDisinfo. (2020). 'EEAS Special Report: Disinformation on the Coronavirus – short assessment of the information environment'. 19 March 2020. https://euvsdisinfo.eu/eeas-special-report-disinformation-on-the-coronavirus-short-assessment-of-the-information-environment/

Evans, R. (2020). 'How Coronavirus Scammers Hide On Facebook And YouTube'. *Bellingcat*, 19 March 2020. https://www.bellingcat.com/news/rest-of-world/2020/03/19/how-coronavirus-scammers-hide-on-facebook-and-youtube/

Facebook. (2018). 'Facts About Content Control on Facebook'. *Facebook Newsroom*, 28 December 2018. https://about.fb.com/news/2018/12/content-review-facts/

Facebook. (2018b). 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar'. 05 November 2018. https://about.fb.com/news/2018/11/myanmar-hria/

Facebook. (2019). Ad Library. https://www.facebook.com/ads/library/

Facebook. (2019b). 'Fact-Checking on Facebook: What Publishers should know'. https://www.facebook.com/help/publisher/182222309230722

Facebook. (2019c). Community Standards. https://www.facebook.com/communitystandards/introduction

Facebook. (2019d). Letter from Rebecca Stimson, Head of Public Policy, UK to Department of Digital, Culture, Media and Sport Committee House of Commons on 29 October 2019. https://www.parliament.uk/documents/commons-committees/culture-media-and-sport/191029%20Rebecca%20Stimson%20Facebook%20to%20Chair%20response%20to%2022%20Oct%20letter.pdf

Facebook. (2019e). Oversight Board Charter. https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf

Facebook. (2020a). 'Working With Industry Partners – Joint industry statement from Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter and YouTube'. 16 March 2020. https://about.fb.com/news/2020/04/coronavirus/#joint-statement

Facebook. (2020b). 'An Update to How We Address Movements and Organizations Tied to Violence'. 19 August 2020. https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/

Facebook and Instagram. (2019). 'Facebook Report on the Implementation of the Code of Practice for Disinformation – Annual Report'. https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019

FactCheck. (2019). 'Our Funding'. https://www.factcheck.org/our-funding/

Falck, B. (2018). 'Providing More Transparency Around Advertising on Twitter', *Twitter Blog*, 28 June 2018. https://blog.twitter.com/official/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html

Faridani, S. (2010). 'Opinion space: a scalable tool for browsing online comments'. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI'10*), *1175-1184. ACM*, April 2010. https://dl.acm.org/doi/10.1145/1753326.1753502

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E. & Benkler, Y. (2017). 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election'. *Berkman Klein Center Research Publication 2017-6*. https://dash.harvard.edu/handle/1/33759251

Fidler, D. P. (2019). 'Disinformation and Disease: Social Media and the Ebola Epidemic in the Democratic Republic of the Congo'. *Council on Foreign Affairs*, 20 August 2019. https://www.cfr.org/blog/disinformation-and-disease-social-media-and-ebola-epidemic-democratic-republic-congo

Fischer, S. (2020). 'Exclusive: Facebook cracks down on political content disguised as local news'. *Axios*, 11 August 2020. https://www.axios.com/facebook-pages-news-exemption-e66d92ce-2abd-4293-b2ad-16cf223e12f1.html

Fitzgibbon, W. (2020). 'Benin investigative journalist jailed in 'absurd' decision'. *ICIJ*, 19 May 2020. https://www.icij.org/blog/2020/05/benin-investigative-journalist-jailed-in-absurd-decision/

Flaxman, S. R., Goel, S. & Rao, J. M. (2016). 'Filter Bubbles, Echo Chambers, and Online News Consumption'. *Public Opinion Quarterly*, 80 (Special issue): 298–320. http://sethrf.com/files/bubbles.pdf

Fletcher, R., & Nielsen, R. K. (2018). 'Are people incidentally exposed to news on social media? A comparative analysis'. *New Media & Society*, 17 August 2017. 20(7), 2450-2468 https://doi.org/10.1177/1461444817724170

Flore, M., Balahur, A., Podavini, A. & Verile, M. (2019). 'Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda'. EUR 29741 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-03320-2, doi:10.2760/919835, JRC116009. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC116009/understanding_citizens_vulnerabilities_to_disinformation.pdf

Foer, F. (2017). 'World Without Mind: The Existential Threat of Big Tech'. Penguin, London.

FOJO: Media Institute. (2018). '#journodefender: Turning trolling against journalists on its head'. October 2018. https://journodefender.org/media/SE_journodefender_public_v1.pdf

Fortune, C. (2018). 'Digitally dissecting atrocities – Amnesty International's open source investigations'. *Amnesty International*, 26 September 2018. https://www.amnesty.org/en/latest/news/2018/09/digitally-dissecting-atrocities-amnesty-internationals-open-source-investigations/

France Fight against Manipulation of Information Law. (2018). 'LOI n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (1)'. *Legifrance*, 23 December 2018. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037847559&categorieLien=id

France 24. (2020). 'Misinformation flood hampers fight for virus vaccine in Africa'. 07 May 2020. https://www.france24.com/en/20200507-misinformation-flood-hampers-fight-for-virus-vaccine-in-africa

François, C. (2019). 'Actors, Behaviors, Content: A Disinformation ABC. Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses', *Transatlantic Working Group*, 20 September 2019. https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf

François, C., Nimmo, B. & Shawn Eib, C. (2019). 'The IRA CopyPasta Campaign'. *Graphika*, October 2019. https://graphika.com/uploads/Graphika%20Report%20-%20CopyPasta.pdf

Frau-Meigs, D. (2019). 'Information Disorders: Risks and Opportunities for Digital Media and Information Literacy?' Media Studies 10, 19 (2019): 11-27 https://hrcak.srce.hr/ojs/index.php/medijske-studije/issue/view/392

Frau-Meigs, D. (in press). 'Addressing the risks of harms caused by disinformation: European vs American approaches to testing the limits of dignity and freedom of expression online'. In Handbook of Communication Rights, Law & Ethics, L. Corredoira (ed), Wiley, forthcoming.

Freedman, D. (2016). 'Divided Britain? We were already divided...' In EU Referendum Analysis 2016: Media, Voters and the Campaign. Jackson, D., Thorsen, E., Wring, D. Loughborough University Center for the Study of Journalism, Culture and Community. http://www.referendumanalysis.eu/

Free Malaysia Today. (2019). 'Malaysia finally scraps Anti-Fake News Act'. 19 December 2019. https://www.freemalaysiatoday.com/category/nation/2019/12/19/malaysia-finally-scraps-anti-fake-news-act/

French Parliament. (2019). 'Lutte contre la haine sur internet - Proposition de loi. n° 1785 , déposé(e) le mercredi 20 mars 2019'. http://www.assemblee-nationale.fr/dyn/15/dossiers/lutte_contre_haine_internet

Friedman, U. (2020). 'The Coronavirus-Denial Movement Now Has a Leader'. *The Atlantic*, 27 March 2020. https://www.theatlantic.com/politics/archive/2020/03/bolsonaro-coronavirus-denial-brazil-trump/608926/

Fries, F. (2018). 'The Role of a Global News Agency In The Era of Big Tech And Fake News'. *FCC*, 10 December 2018. https://www.fcchk.org/event/club-lunch-the-role-of-a-global-news-agency-in-the-era-of-big-tech-and-fake-news/

Full Fact. (2018). 'Tackling misinformation in an open society'. https://fullfact.org/media/uploads/full_fact_tackling_misinformation_in_an_open_society.pdf

Full Fact (2019). 'Report on the Facebook Third Party Fact Checking programme'. https://fullfact.org/media/uploads/tpfc-q1q2-2019.pdf

Funke. D. (2018). 'Automated fact-checking has come a long way. But it still faces significant challenges'. *Poynter*, 4 April, 2018. https://www.poynter.org/news/automated-fact-checking-has-come-long-way-it-still-faces-significant-challenges

Funke, D. (2019). ''I spent almost a month on a floor': What it's like to be imprisoned on false news charges'. *Poynter*, 22 January 2019. https://www.poynter.org/fact-checking/2019/i-spent-almost-a-month-on-a-floor-what-its-like-to-be-imprisoned-on-false-news-charges/

Funke, D. & Benkelman, S. (2019). 'Factually: Games to teach media literacy'. American Press Institute, 18 July 2019. https://www.americanpressinstitute.org/fact-checking-project/factually-newsletter/factually-games-to-teach-media-literacy/

Funke, D. & Mantzarlis, A. (2018b). 'Here's what to expect from fact-checking in 2019'. *Poynter*, 18 December 2018. https://www.poynter.org/fact-checking/2018/heres-what-to-expect-from-fact-checking-in-2019/

Funke, D. & Mantzarlis, A. (2018a). 'We asked 19 fact-checkers what they think of their partnership with Facebook. Here's what they told us'. *Poynter*, 14 December 2018. https://www.poynter.org/fact-checking/2018/we-asked-19-fact-checkers-what-they-think-of-their-partnership-with-facebook-heres-what-they-told-us/

Gadde, V. (2020). 'Protecting and supporting journalists during COVID-19'. *Twitter*, 24 March 2020. https://blog.twitter.com/en_us/topics/company/2020/giving-back-covid-19.html

Gadde, V. & Derella, M. (2020). 'An update on our continuity strategy during COVID-19'. *Twitter Blog*, 16 March 2020. https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html

Galbraith, J. K. (1971). 'A contemporary guide to economics, peace and laughter. Essays edited by Andrea D. Williams'. Chapter 3: How Keynes Came to America, Quote Page 50, Houghton Mifflin Company, Boston, Massachusetts.

Galtung, J. & Ruge, M.H. (1965). 'The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers'. *Journal of Peace Research*, 2(1), 64-90. https://www.semanticscholar.org/paper/The-Structure-of-Foreign-News-The-Presentation-of-Galtung-Ruge/b3b929df1fd2aa3ea6ddd7b44448fd409e48ea0a

Garside, J. (2020). 'Rappler editor Maria Ressa: 'They could shut us down tomorrow''. *The Guardian*, 26 February 2020. https://www.theguardian.com/world/2020/feb/26/rappler-editor-maria-ressa-journalist-they-could-shut-us-down-tomorrow-philippines-fake-news

Gaw, F. (2020). 'Digital disinformation is as potent as a virus during a pandemic'. *Rappler*, 20 March 2020. https://www.rappler.com/technology/features/255224-digital-disinformation-fake-news-coronavirus

Gentzkow, M. & Shapiro, J. M. (2011). 'Ideological segregation online and offline." *Quarterly Journal of Economics*, 126 (4): 1799–1839. (DOI): 10.3386/w15916 https://www.nber.org/papers/w15916

German BMJV. (2020a). 'Gesetzespaket gegen Rechtsextremismus und Hasskriminalität, 2020'. *German Federal Ministry of Justice and Consumer Protection*. 19 February, 2020. https://www.bmjv.de/SharedDocs/Artikel/DE/2020/021920_Kabinett_Bekaempfung_Rechtsextremismus_Hasskriminalitaet.html

German BMJV. (2020b). 'Gesetz zur Bekämpfung des Rechtsextremismus und der Hasskriminalität, 2020'. German Federal Ministry of Justice and Consumer Protection. https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_NetzDGAendG.pdf?__blob=publicationFile&v=3

German NetzDG. (2017). 'Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG)'. *Deutscher Bundestag*. https://dipbt.bundestag.de/extrakt/ba/WP18/815/81582.html

German NetzDG English translation. (2017). 'Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG)' German Law Archive, 1 October 2017. https://germanlawarchive.iuscomp.org/?p=1245

Gettleman, J., Goel, V. & Abi-Habib, M. (2019). 'India Adopts the Tactic of Authoritarians: Shutting Down the Internet'. *The New York Times*, 17 December 2019. https://www.nytimes.com/2019/12/17/world/asia/india-internet-modi-protests.html

Giglietto, F., Iannelli, L., Rossi, L. & Valeriani, A. (2016). 'Fakes, News and the Election: A New Taxonomy for the Study of Misleading Information within the Hybrid Media System'. Convegno AssoComPol 2016. SSRN, https://ssrn.com/abstract=2878774

GIJN Staff. (2019). 'Full Text: Maria Ressa's Keynote Speech for #GIJC19'. 08 October 2019. https://gijn.org/2019/10/08/full-text-maria-ressas-keynote-speech-for-gijc19/

Gilbert, B. (2019). 'Facebook refuses to fact-check political ads, and it's infuriating employees and lawmakers. Here's why the issue continues to dog the company.' *Business Insider*, 14 December 2019. https://www.businessinsider.com/facebook-political-ads-fact-check-policy-explained-2019-11?r=US&IR=T

Gillespie, T. (2017). 'The platform metaphor, revisited'. *Hig Science Blog, Institut fur Internet und Gesellschaft*, 24 August 2017. https://www.hiig.de/en/the-platform-metaphor-revisited/

Glazer, E. (2019). 'Facebook Weighs Steps to Curb Narrowly Targeted Political Ads'. *The Wall Street Journal*, 21 November 2019. https://www.wsj.com/articles/facebook-discussing-potential-changes-to-political-ad-policy-11574352887?redirect=amp

Gleicher, N. (2018a). 'Coordinated Inauthentic Behavior Explained'. *Facebook*, 6 December 2018. https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

Gleicher, N. (2018b). 'How We Work With Our Partners to Combat Information Operations'. *Facebook Newsroom*, 13 November 2018. https://about.fb.com/news/2018/11/last-weeks-takedowns/#working-with-partners

Gleicher, N. (2019). 'How We Respond to Inauthentic Behavior on Our Platforms: Policy Update'. *Facebook Newsroom*, 21 October 2019. https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update/

Gleicher, N. (2020). 'Removing Coordinated Inauthentic Behavior From Russia, Iran, Vietnam and Myanmar'. *Facebook*, 12 February, 2020. https://about.fb.com/news/2020/02/removing-coordinated-inauthentic-behavior/

Global Disinformation Index. (2019). 'The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech?' September 2019. https://disinformationindex.org/wp-content/uploads/2019/09/GDI_Ad-tech_Report_Screen_AW16.pdf

Goel, V., Deep Singh, K. & Yasir, S. (2019). 'India Shut Down Kashmir's Internet Access. Now, 'We Cannot Do Anything.'. *The New York Times*, 14 August 2019. https://www.nytimes.com/2019/08/14/technology/india-kashmir-internet.html

Goggin, B. & Tenbarge, K. (2019). ''Like you've been fired from your job': YouTubers have lost thousands of dollars after their channels were mistakenly demonetized for months'. *Business Insider*, 24 August 2019. https://www.businessinsider.com/youtubers-entire-channels-can-get-mistakenly-demonetized-for-months-2019-8?r=US&IR=T

Goldshlager, K. & Watson, O. (2020). 'Launching a $1M Grant Program to Support Fact-Checkers Amid COVID-19'. *Facebook Journalism Project*, 30 April 2020. https://www.facebook.com/journalismproject/coronavirus-grants-fact-checking

Goldzweig, R. (2020). 'It is time tech companies act on election-time disinformation'. *Al Jazeera*, 23 May 2020. https://www.aljazeera.com/indepth/opinion/time-tech-companies-act-election-time-disinformation-200520135809708.html

Google (2019). Political Advertising on Google. https://transparencyreport.google.com/political-ads/home

Google (2019). YouTube Community Guidelines Enforcement. https://transparencyreport.google.com/youtube-policy/removals

Google and YouTube (2019). 'EC EU Code of Practice on Disinformation – Google Annual Report', https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019

Gorwa, R., Binns, R. & Katzenbach, C. (2020). 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', *Big Data & Society*, January - June 2020, 1-15. https://doi.org/10.1177/2053951719897945

Gottfried, J. & Greco, E. (2018). 'Younger Americans are better than older Americans at telling factual news statements from opinions'. *Pew Research Center*, 23 October 2018. http://www.pewresearch.org/fact-tank/2018/10/23/younger-americans-are-better-than-older-americans-at-telling-factual-news-statements-from-opinions/,

Gottfried, J., Barthel, M. & Mitchell, A. (2017). 'Trump, Clinton Voters Divided in Their Main Source for Election News'. *Pew Research Centre*, 18 January 2017. https://www.journalism.org/2017/01/18/trump-clinton-voters-divided-in-their-main-source-for-election-news/

Government of Canada. (2019a). 'Response to Canadian Parliamentary Standing Committee on Access to Information, Privacy and Ethics Report. Democracy under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly'. https://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-17/response-8512-421-502

Government of Canada. (2019b). 'Online Disinformation'. https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html

Graells-Garrido, E., Lalmas, M. & Baeza-Yates, R. (2016). 'Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles'. *In Proceedings of the 21st International Conference on Intelligent User Interfaces*, March 2016 https://doi.org/10.1145/2856767.2856776

Grau, M. (2020). 'New WhatsApp chatbot unleashes power of worldwide fact-checking organizations to fight COVID-19 misinformation on the platform'. *Poynter*, 04 May 2020. https://www.poynter.org/fact-checking/2020/poynters-international-fact-checking-network-launches-whatsapp-chatbot-to-fight-covid-19-misinformation-leveraging-database-of-more-than-4000-hoaxes/

Graves, L. (2013). 'Deciding What's True: Fact-Checking Journalism and the New Ecology of News'. Columbia University. Philosophy PhD dissertation. https://core.ac.uk/download/pdf/161442732.pdf

Graves, L. (2018). 'FACTSHEET: Understanding the Promise and Limits of Automated Fact-Checking'. *Reuters*. http://www.digitalnewsreport.org/publications/2018/factsheet-understanding-promise-limits-automated-fact-checking/

Graves, L, & Cherubini, F. (2016). 'The Rise of Fact-Checking Sites in Europe'. *Reuters*, http://www.digitalnewsreport.org/publications/2016/rise-fact-checking-sites-europe/

Green, V. (2019). 'Our staff was clear: Facebook's money isn't worth it'. https://twitter.com/vinnysgreen/status/1187135050761920512

Green, V. & Mikkelson, D. (2019). 'A Message to Our Community Regarding the Facebook Fact-Checking Partnership'. *Snopes*, 1 February 2019. https://www.snopes.com/blog/2019/02/01/snopes-fb-partnership-ends/

Gregory, S. (2019). 'Deepfakes and Synthetic Media: Updated Survey of Solutions against Malicious Usages'. *Witness*, June 2019. https://blog.witness.org/2019/06/deepfakes-synthetic-media-updated-survey-solutions-malicious-usages/

Gregory, S. & French, E. (2019). 'How do we work together to detect AI-manipulated media?'. *Witness Media Lab*. https://lab.witness.org/projects/osint-digital-forensics/

Grigoryan, A. (2019). 'Initiatives to Counter Fake News: Russia'. *Library on Congress*, April 2019. https://www.loc.gov/law/help/fake-news/russia.php#_ftn17

Grose, C. R. & Husser, J.A. (2008). 'The Valence Advantage of Presidential Persuasion: Do Presidential Candidates Use Oratory to Persuade Citizens to Vote Contrary to Ideological Preferences?' 03 April 2008. https://ssrn.com/abstract=1119444 or http://dx.doi.org/10.2139/ssrn.1119444

Grossman, D. & Schickler, R. (2019). 'Facebook took down our fact-check on medically necessary abortions. That's dangerous.' *The Washington Post*, 15 September 2019. https://www.washingtonpost.com/opinions/2019/09/15/facebook-took-down-our-fact-check-medically-necessary-abortions-thats-dangerous/

Guadagno, R. E., & Guttieri, K. (2019). 'Fake News and Information Warfare: An Examination of the Political and Psychological Processes From the Digital Sphere to the Real World'. In *Handbook of Research on Deception, Fake News, and Misinformation Online* (pp. 167-191). IGI Global

Guess, A., Nagler, J. and Tucker, J. (2019). 'Less than you think: Prevalence and predictors of fake news dissemination on Facebook'. *Science Advances*, Vol. 5 (1), p.eaau4586. https://advances.sciencemag.org/content/5/1/eaau4586

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018a). 'Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think'. Knight Foundation White Paper. https://www.researchgate.net/publication/330144926_Avoiding_the_echo_chamber_about_echo_chambers_Why_selective_exposure_to_like-minded_political_news_is_less_prevalent_than_you_think

Guess, A., Nyhan, B., Reifler, J. (2018b). "Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 US. Presidential Campaign". http://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Gupta, S. (2020). 'Journalism in the time of corona: This is the biggest story of our lives'. *The Print*, 21 March 2020. https://theprint.in/national-interest/journalism-in-the-time-of-corona-this-is-the-biggest-story-of-our-lives/385057/?amp&__twitter_impression=true

Haffajee, F. & Davies, M. (2017). 'Ferial Haffajee: The Gupta Fake News Factory And Me'. *Huffington Post*, 06 June 2017. https://www.huffingtonpost.co.uk/2017/06/05/ferial-haffajee-the-gupta-fake-news-factory-and-me_a_22126282/

Haffajee, F. (2019). '#GuptaLeaks wins Global Shining Light investigative journalism award'. *Daily Maverick*, 29 September 2019. https://www.dailymaverick.co.za/article/2019-09-29-guptaleaks-wins-global-shining-light-investigative-journalism-award/

Halon, Y. (2020). 'Zuckerberg knocks Twitter for fact-checking Trump, says private companies shouldn't be 'the arbiter of truth'. *Fox News*, 28 May 2020. https://www.foxnews.com/media/facebook-mark-zuckerberg-twitter-fact-checking-trump

Hamborg, F., Donnay, K. & Gipp, B. (2018). 'Automated identification of media bias in news articles: an interdisciplinary literature review'. 16 November 2018. Int J Digit Libr 20, 391–415 (2019). https://doi.org/10.1007/s00799-018-0261-y

Hanly, K. (2018). 'Op-Ed: Linguist George Lakoff explains how Trump tactics work'. Digital Journal, 17 November 2018. http://www.digitaljournal.com/news/politics/op-ed-linguist-george-lakoff-explains-how-trump-tactics-work/article/537117

Hansen, F. S. (2017). 'The weaponization of information'. *DIIS*, 14 December 2017. https://www.diis.dk/en/research/the-weaponization-of-information

Hao, K. (2019). 'This is how AI bias really happens—and why it's so hard to fix'. *MIT Technology Review*, 04 February 2019. https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Hanusch, F. (2017). 'Web analytics and the functional differentiation of journalism cultures: Individual, organizational and platform-specific influences on newswork'. Information, Communication & Society, 20(10), 1571-1586. https://www.tandfonline.com/doi/full/10.1080/1369118X.2016.1241294

Harding McGill, M. & Daly, K. (2020). 'New calls to curtail tech's targeted political advertising'. *Axios*, 28 May 2020. https://www.axios.com/new-calls-to-curtail-techs-targeted-political-advertising-9dcd7d8a-4d27-4846-92a1-6f4aab42d7b2.html

Harris, B. (2019). 'An Update on Building a Global Oversight Board'. *Facebook Newsroom*, 12 December 2019. https://about.fb.com/news/2019/12/oversight-board-update/

Harvey, D. (2019). 'Helping you find reliable public health information on Twitter'. *Twitter Blog*, 10 May 2019. https://blog.twitter.com/en_us/topics/company/2019/helping-you-find-reliable-public-health-information-on-twitter.html

Harvey, D. & Roth, Y. (2018). 'An update on our elections integrity work'. *Twitter*, 1 October 2018. https://blog.twitter.com/official/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html

Hatmaker, T. (2020). 'Jack Dorsey explains why Twitter fact-checked Trump's false voting claims'. *Techcrunch*, 28 May 2020. https://techcrunch.com/2020/05/27/twitter-vs-trump-fact-checking-dorsey/

Hazard Owen, L. (2019). 'Full Fact has been fact-checking Facebook posts for six months. Here's what they think needs to change'. *Nieman Lab*, 29 July 2019. https://www.niemanlab.org/2019/07/full-fact-has-been-fact-checking-facebook-posts-for-six-months-heres-what-they-think-needs-to-change/

Henderson, A. (2020). 'NYT slammed for 'terminal both-sides-ism' after reporting on Trump's household disinfectant suggestion'. *AlterNet*, 24 April 2020. https://www.alternet.org/2020/04/nyt-slammed-for-terminal-both-sides-ism-after-reporting-on-trumps-household-disinfectant-suggestion/

Henley, J. (2020). 'How Finland starts its fight against fake news in primary schools'. *The Guardian*, 29 January 2020. https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news

Hern, A. (2019a). 'Facebook's only Dutch factchecker quits over political ad exemption'. *The Guardian*, 27 November 2019. https://www.theguardian.com/technology/2019/nov/27/facebook-only-dutch-factchecker-quits-over-political-ad-exemption

Hern, A. (2019b) 'Facebook fact checkers did not know they could vet adverts'. *The Guardian*, 26 October 2019. https://www.theguardian.com/technology/2019/oct/26/facebook-fact-checkers-paid-adverts-misinformation-mark-zuckerberg-congress

Hern, A. (2020). 'YouTube ads of 100 top brands fund climate misinformation – study'. *The Guardian*, 16 January 2020. https://www.theguardian.com/technology/2020/jan/16/youtube-ads-of-100-top-brands-fund-climate-misinformation-study

Hogan, L. (2018). 'Myanmar groups criticise Zuckerberg's response to hate speech on Facebook'. *The Guardian*, 05 April 2018. https://www.theguardian.com/technology/2018/apr/06/myanmar-facebook-criticise-mark-zuckerberg-response-hate-speech-spread

Hoggins, T. (2019). 'Google and Facebook's moderators are superheroes, but they need more protection'. *The Telegraph*, 17 December 2019. https://www.telegraph.co.uk/technology/2019/12/17/google-facebooks-moderators-superheroes-need-protection/

Hollowood, E. & Mostrous, A. (2020). 'Fake news in the time of C-19: From miraculous cures to paranoid conspiracies, our investigation reveals how misinformation about coronavirus is going viral at a disturbing rate'. *Tortoise*, 23 March 2020. https://members.tortoisemedia.com/2020/03/23/the-infodemic-fake-news-coronavirus/content.html

Hollyfield, A. (2013). 'PolitiFact to launch PunditFact, checking pundits and media figures'. *Politifact*, 10 October 2013. https://www.politifact.com/article/2013/oct/10/politifact-launch-punditfact-checking-pundits-and-/

Horwitz, J. & Seetharaman, D. (2020). 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive'. *The Wall Street Journal*, 26 May 2020. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?utm_source=Daily+Lab+email+list&utm_campaign=9e6150a2bf-dailylabemail3&utm_medium=email&utm_term=0_d68264fd5e-9e6150a2bf-396529883

Houngbadji, C. S. (2020). 'Bénin: après Ignace Sossou, un autre journaliste fait les frais du code du numérique'. *Benin Web TV*, 07 January 2020. https://beninwebtv.com/2020/01/benin-apres-ignace-sossou-un-autre-journaliste-fait-les-frais-du-code-du-numerique/

House of Lords. Select Committee on Democracy and Digital Technologies. Digital Technology and the Resurrection of Trust. HL Paper 77. 29 June 2020. https://publications.parliament.uk/pa/ld5801/ldselect/lddemdigi/77/77.pdf

Howard, P. & Bell, E. (2020). 'Disinformation in 2020, from "Plandemic" to Bill Gates to "Obamagate". Interview with Recode Decode's Kara Swisher, 27 May 2020. https://www.podchaser.com/podcasts/recode-decode-100800/episodes/phil-howard-and-emily-bell-dis-61618752

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J. & François, C. (2018). 'The IRA, Social Media and Political Polarization in the United States, 2012-2018'. *Computational Propaganda Research Project*. https://comprop.oii.ox.ac.uk/research/ira-political-polarization/

Human Rights Watch. (2018a). 'Kazakhstan: Criminal Probe of Media Outlets'. 06 April 2018. https://www.hrw.org/news/2018/04/06/kazakhstan-criminal-probe-media-outlets#

Human Rights Watch. (2018b). 'Germany: Flawed Social Media Law'. 14 February 2018. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Humprecht, E., Esser, F. & Van Aelst P. (2020). 'Resilience to Online Disinformation: A Framework for Cross-National Comparative Research'. 24 January 2020. https://doi.org/10.1177/1940161219900126

IAMAI. (2019). 'Voluntary Code of Ethics for the General Elections 2019'. http://164.100.117.97/WriteReadData/userfiles/Voluntary%20Code%20of%20Ethics%20for%20the%20G.E.%202019.pdf.

IFCN. (2019a). The International Fact-Checking Network, *Poynter*. https://www.poynter.org/ifcn/

IFCN. (2019c). 'The commitments of the code of principles'. https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles. Visited in December 2019.

IFCN. (2019d). 'About the International Fact-Checking Network'. https://www.poynter.org/about-the-international-fact-checking-network/. Visited in December 2019.

IFCN. (2020a). 'IFCN receives $1 million from WhatsApp to support fact-checkers on the coronavirus battlefront'. *Poynter*, 18 March 2020. https://www.poynter.org/fact-checking/2020/ifcn-receives-1-million-from-whatsapp-to-support-fact-checkers-on-the-coronavirus-battlefront/

IFCN. (2020b). 'Flash grants of up to $50K are now available for fact-checkers fighting coronavirus misinformation'. *Poynter*, 17 March 2020. https://www.poynter.org/fact-checking/2020/flash-grants-of-up-to-50k-are-now-available-for-fact-checkers-fighting-coronavirus-misinformation/

IFCN. (2020c). 'The Coronavirus Fact-Checking Grants will support 13 projects on its 1st round: More than half a million dollars is being distributed'. *Poynter*, 02 April 2020. https://www.poynter.org/fact-checking/2020/the-coronavirus-fact-checking-grants-will-support-13-projects-on-its-1st-round-more-than-half-a-million-dollars-is-being-distributed/

IFCN. (2020d). 'Verified signatories of the IFCN code of principles'. https://ifcncodeofprinciples.poynter.org/signatories. Visited on 6 August 2020.

Indian (MeitY) Government (2018). 'The Information Technology [Intermediaries Guidelines (Amendment) Rules] 2018' https://www.meity.gov.in/writereaddata/files/Draft_Intermediary_Amendment_24122018.pdf

Indian Ministry of Communications. (2017). 'Temporary Suspension of Telecom Services Rules'. https://dot.gov.in/sites/default/files/Suspension%20Rules.pdf?download=1

Ingber, S. (2019). 'Students in Ukraine Learn How To Spot Fake Stories, Propaganda And Hate Speech'. *NPR*, 22 March 2019. https://www.npr.org/2019/03/22/705809811/students-in-ukraine-learn-how-to-spot-fake-stories-propaganda-and-hate-speech?t=1581335026341

Ingram, M. (2018). 'Facebook slammed by UN for its role in Myanmar genocide'. *CJR*, 08 November 2018. https://www.cjr.org/the_media_today/facebook-un-myanmar-genocide.php

Ingram, M. (2019). 'YouTube is all over the map when it comes to offensive content'. *CJR*, 6 June 2019. https://www.cjr.org/the_media_today/youtube-maza-nazis.php

Ireland House of the Oireachtas. (2019). *International Grand Committee on Disinformation and 'Fake News' Dublin, Ireland – Wednesday 6th and Thursday 7th November 2019*. Press Release, 25 October 2019. https://www.oireachtas.ie/en/press-centre/press-releases/20191025-

international-grand-committee-on-disinformation-and-fake-news-dublin-ireland-wednesday-6th-and-thursday-7th-november-2019/

Ireland IDG, (2019). 'Government of Ireland, Interdepartmental Group on Security of Ireland's Electoral Process and Disinformation – Progress Report'. https://assets.gov.ie/39188/8c7b6bc1d0d046be915963abfe427e90.pdf

Ireton C. & Posetti, J., eds. (2018). 'Journalism, Fake News' & Disinformation: A Handbook for Journalism Education and Training', pp 21-22. *UNESCO*. https://en.unesco.org/fightfakenews

Irish Department of the Taoiseach. (2019). 'Proposal to Regulate Transparency of Online Political Advertising'. Press Release, 05 November 2019. https://www.gov.ie/en/news/9b96ef-proposal-to-regulate-transparency-of-online-political-advertising/

Irish IDG. (2019). 'Government of Ireland - Interdepartmental Group on Security of Ireland's Electoral Process and Disinformation – Progress Report'. November, 2019. https://assets.gov.ie/39188/8c7b6bc1d0d046be915963abfe427e90.pdf

ISD. (2019). '2019 EU Elections Information Operations Analysis'. https://www.isdglobal.org/wp-content/uploads/2019/05/ISD-EU-Elections-Computational-Propaganda-Analysis_May-24.docx.pdf

ISD. (2020a). 'Covid-19 Disinformation Briefing No.1'. 27 March 2020. https://www.isdglobal.org/wp-content/uploads/2020/03/COVID-19-Briefing-Institute-for-Strategic-Dialogue-27th-March-2020.pdf

ISD. (2020b). 'Covid-19 Disinformation Briefing No.2'. 09 April 2020. https://www.isdglobal.org/wp-content/uploads/2020/04/Covid-19-Briefing-PDF.pdf

Italian Ministry of Education. (2017). 'Scuola, Boldrini e Fedeli presentano decalogo anti-bufale Il progetto riguarderà 4,2 milioni di ragazzi'. Press Release, 31 October 2017. https://www.miur.gov.it/web/guest/-/scuola-boldrini-e-fedeli-presentano-decalogo-anti-bufale-il-progetto-riguardera-4-2-milioni-di-ragazzi

Italian AGCOM. (2018). 'Linee guida per la parità di accesso alle piattaforme online durante la campagna elettorale per le elezioni politiche 2018'. https://www.agcom.it/documents/10179/9478149/Documento+generico+01-02-2018/45429524-3f31-4195-bf46-4f2863af0ff6?version=1.0

Itimu, K. (2019). 'Judgement on the Computer Misuse and Cybercrimes Law to be Passed Next Year'. *Techweez*, 24 October 2019. https://techweez.com/2019/10/24/judgement-on-the-computer-misuse-and-cybercrimes-law-to-be-passed-next-year/

Jack, C. (2017). 'Lexicon of Lies: Terms for Problematic Information'. *Data & Society Research*, 09 August 2017. https://datasociety.net/output/lexicon-of-lies/

Jahangir, R. (2020). 'Desi totkas and fake news — a guide to surviving the Covid-19 'infodemic'. DAWN, 28 March 2020. https://www.dawn.com/news/1544256/desi-totkas-and-fake-news-a-guide-to-surviving-the-covid-19-infodemic

Jamison, A., Broniatowski, D., Dredze, M., Wood-Doughty, Z., Khan, D. & Crouse Quinn, S. (2019). 'Vaccine-related advertising in the Facebook Ad Archive'. *Science Direct*, Vol 38 :3, 16 January 2020. https://www.sciencedirect.com/science/article/pii/S0264410X1931446X?via%3Dihub

Jankowicz, N. (2019). 'Ukraine's Election Is an All-Out Disinformation Battle'. *The Atlantic*, 17 April 2019. https://www.theatlantic.com/international/archive/2019/04/russia-disinformation-ukraine-election/587179/

Japanese Ministry of Internal Affairs and Communications. (2018).「プラットフォームサービスに関する研究会における検討アジェンダ（案）」に対する提案募集. Press Release, 19 October 2018. https://www.soumu.go.jp/menu_news/s-news/01kiban18_01000052.html

Jeangène Vilmer, J.-B., Escorcia, A., Guillaume, M., & Herrera, J. (2018). 'Information Manipulation: A Challenge for Our Democracies'. A report by the Policy Planning Staff (CAPS Ministry for Europe and Foreign Affairs) and the Institute for Strategic Research (IRSEM, Ministry for the Armed Forces). *Resource Centre on Media Freedom in Europe*, August 2018. https://www.rcmediafreedom.eu/Publications/Reports/Information-Manipulation-A-Challenge-for-Our-Democracies

Jiahao, W. (2019). '防制不實訊息 臉書LINE等5大業者帶頭自律'. *CNA*, 21 June 2019. https://www.cna.com.tw/news/firstnews/201906210183.aspx

Jiji. (2019). 'Japanese panel wants to establish team to fight fake news, with help from U.S. tech giants'. *The Japan Times*, 30 November 2019. https://www.japantimes.co.jp/news/2019/11/30/business/japan-fake-news-gafa/#.XkEz40HgouV

Jin, K-X. (2020). 'Keeping People Safe and Informed About the Coronavirus'. *Facebook Newsroom*, 04 May 2020. https://about.fb.com/news/2020/05/coronavirus/#joint-statement

Johnson, H. M. & Seifert, C. M. (1994). 'Sources of the continued influence effect: When misinformation in memory affects later inferences'. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Kafka, P. 'Rupert Murdoch wanted Mark Zuckerberg to pay him for news stories — and now Facebook is going to do just that'. *Vox*, 24 October 2019. https://www.vox.com/recode/2019/10/24/20929919/facebook-zuckerberg-murdoch-news-publishers-pay-content

Kahn, B. (2019). 'This Fake Green New Deal Ad Perfectly Illustrates Facebook's Bullshit Political Ad Policy [Updated]. *Gizmodo*, 25 October 2019. https://earther.gizmodo.com/this-fake-green-new-deal-ad-perfectly-illustrates-faceb-1839364467

Kajimoto, M. (2018). 'In East and Southeast Asia, misinformation is a visible and growing concern'. *Poynter*, 14 March 2018. https://www.poynter.org/fact-checking/2018/in-east-and-southeast-asia-misinformation-is-a-visible-and-growing-concern/

Kajimoto, M., Stanley S. Editors. (2019). 'Information Disorder in Asia and the Pacific'. *The University of Hong-Kong, Journalism and Media Studies Center*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3134581

Kalogeropoulos, A., Cherubini, F., & Newman, N. (2016). 'The future of online news video'. *Digital News Project*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/files/The%2520Future%2520of%2520Online%2520News%2520Video.pdf

Kang C and Frenkel S (2020) "Facebook removes Trump campaign's misleading coronavirus video" in *The New York Times*. August 5th, 2020. https://www.nytimes.com/2020/08/05/technology/trump-facebook-coronavirus-video.html

Kao, J. (2020). 'How China Built a Twitter Propaganda Machine Then Let It Loose on Coronavirus'. *ProPublica*, 26 March 2020. https://www.propublica.org/article/how-china-built-a-twitter-propaganda-machine-then-let-it-loose-on-coronavirus

Karanbir Gurung, S. (2019). 'Defence ministry approves information warfare branch for Indian army'. *The Economic Times*, 09 march 2019. https://economictimes.indiatimes.com/news/defence/defence-ministry-approves-information-warfare-branch-for-indian-army/articleshow/68329797.cms

Karimi, N. & Gambrell, J. (2020). 'In Iran, false belief a poison fights virus kills hundreds'. *AP News*, 27 March 2020. https://apnews.com/6e04783f95139b5f87a5febe28d72015

Kaur, K., Nair, S., Kwok, Y., Kajimoto, M, Chua, Y. T., Labiste, M. D., Soon, C., Jo, H., Lin, L., Le, T. T. & Kruger, A. (2018). Chapter on India in 'Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam'. 10 October 2018. http://dx.doi.org/10.2139/ssrn.3134581

Kaye, D. (2018). 'Report of the special rapporteur on promotion and protection of the right to freedom of opinion and expression'. *Freedex*. https://freedex.org/a-human-rights-approach-to-platform-content-regulation/

Kaye, D. (2020a). 'Disease pandemics and the freedom of opinion and expression'. 23 April 2020. https://freedex.org/wp-content/blogs.dir/2015/files/2020/04/A_HRC_44_49_AdvanceEditedVersion.pdf

Kaye, D. (2020b). Transcript of interview with Recode Decode's Kara Swisher. https://docs.google.com/document/d/1zYuYMTmvoHu_dgajRTSORZpkY1DhiN8Kb9KytCgVU7g/edit Recode Decode episode: S1, E480, 14 February 2020.

Kazakhstan Penal Code. (2014). https://www.unodc.org/res/cld/document/penal-code_html/New_penal_code.pdf

Keane, J. (2018). 'Post-truth politics and why the antidote isn't simply 'fact-checking' and truth'. *The Conversation*, 23 March 2018. https://theconversation.com/post-truth-politics-and-why-the-antidote-isnt-simply-fact-checking-and-truth-87364

Keller, I. (2019). 'A school of bitter experience: how Kazakhstan's media regulations restrict journalist freedom'. *Open Democracy*, 6 June 2019. https://www.opendemocracy.net/en/odr/press-freedom-kazakhstan-en/

Keller, F. B., Schoch, D., Stier, S & Yang, J. H. (2019). "Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign." Political Communication (2019): 1-25. https://www.tandfonline.com/doi/full/10.1080/10584609.2019.1661888

Kelly, J. & François, C. (2018). 'This is what filter bubbles actually look like'. *MIT Technology Review*. https://www.technologyreview.com/s/611807/this-is-what-filter-bubbles-actually-look-like/

Kelly, M. (2020a). 'The World Health Organization has joined TikTok to fight coronavirus misinformation'. *The Verge*, 28 February 2020. https://www.theverge.com/2020/2/28/21158276/coronavirus-covid19-tiktok-who-world-health-organization-protection

Kelly, M. (2020b). 'Democrats want to restrict political ad targeting ahead of the 2020 election'. *The Verge*, 26 May 2020. https://www.theverge.com/2020/5/26/21271074/facebook-google-microtargeting-political-ads-ban-anna-eshoo

Kelly Garrett, R. (2009). 'Echo chambers online?: Politically motivated selective exposure among Internet news users'. *Journal of Computer-Mediated Communication*, 14(2), 265-285, 30 March 2019. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1083-6101.2009.01440.x

Kenya Computer Misuse and Cybercrimes Act. (2018). http://kenyalaw.org/kl/fileadmin/pdfdownloads/Acts/ComputerMisuseandCybercrimesActNo5of2018.pdf

Kessler, G. (2017). 'Washington Post: About the Fact Checker'. https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker/

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B. & Potthast, M. (2019). 'SemEval-2019 Task 4: Hyperpartisan News Detection'. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, June 2019. (ππ. 829-839). ΔΟΙ: ¨10.18653/ω1/Σ19-2145 https://www.aclweb.org/anthology/S19-2145

King, G. & Persily, N. (2020). 'Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One'. *Social Science One*, 13 February 2020. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

Khidhir, S. (2019). 'Indonesia has a fake news problem'. *The Asean Post*, 19 November 2019. https://theaseanpost.com/article/indonesia-has-fake-news-problem

Kleinman, Z. (2016). 'Fury over Facebook 'Napalm girl' censorship'. *BBC*, 09 September 2016. https://www.bbc.co.uk/news/technology-37318031#:~:text=Facebook%20said%20it%20has%20to,and%20before%20it%20had%20responded

Knight Foundation. (2018). 'In the internet we trust: the impact of engaging with news articles'. https://knightfoundation.org/reports/in-the-internet-we-trust-the-impact-of-engaging-with-news-articles/

Knockel, J. & Xiong, R. (2019). '(Can't) Picture This 2: An Analysis of WeChat's Realtime Image Filtering in Chats'. *The Citizen Lab*, 15 July 2019. https://citizenlab.ca/2019/07/cant-picture-this-2-an-analysis-of-wechats-realtime-image-filtering-in-chats/

Kongkea, B. R. (2019a). 'Man charged over YouTube fake news'. *Khmer Times*, 02 April 2019. https://www.khmertimeskh.com/592313/man-charged-over-airing-of-fake-news-on-youtube/

Kongkea, B. R. (2019b). 'Former monk convicted over fake news'. *Khmer Times*, 21 October 2019. https://www.khmertimeskh.com/50652677/former-monk-convicted-over-fake-news

Kovach, B. & Rosenstiel, T. (2001). 'The Elements of Journalism, Crown publishers'. New York.

Kozlowska, H. (2019). 'TikTok banned political ads—but pro-Trump content is thriving and misleading teens'. *Quartz*, 24 October 2019. https://qz.com/1731170/pro-trump-videos-are-thriving-on-tiktok/

Kroes, R. (2012). 'The power of rhetoric and the rhetoric of power: Exploring a tension within the Obama presidency'. *European journal of American studies*, *7*(7-2). https://journals.openedition.org/ejas/9578

Kuczerawy, A. (2019). 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?'. Forthcoming in: "Disinformation and Digital Media as a Challenge for Democracy" European Integration and Democracy Series, 6. https://ssrn.com/abstract=3453732

Kutty, S. (2018). 'Fake News - Jail and fine for spreading false news'. *Oman Daily Observer*, 25 May 2018. https://www.omanobserver.om/jail-and-fine-for-spreading-false-news/

La Cour, C. (2019). 'Governments Countering Disinformation: The Case of Italy'. *Disinfo Portal*, 20 November 2019. https://disinfoportal.org/governments-countering-disinformation-the-case-of-italy/

Lacy, L. & Rosenstiel, T. (2015). 'Defining and measuring quality journalism'. *Rutgers*, March 2015. https://www.issuelab.org/resources/31212/31212.pdf

Lamb, K. (2018). 'Cambodia 'fake news' crackdown prompts fears over press freedom'. *The Guardian*, *6* July 2018. https://www.theguardian.com/world/2018/jul/06/cambodia-fake-news-crackdown-prompts-fears-over-press-freedom?CMP=share_btn_tw

Lapowsky, I. (2018). 'Inside the Research Lab Teaching Facebook About Its Trolls'. *Wired*, 15 August 2018. https://www.wired.com/story/facebook-enlists-dfrlab-track-trolls/

Larson, H. J. (2018). 'The biggest pandemic risk? Viral misinformation'. *Natureresearch*, 16 October 2018. Nature 562, 309 (2018). doi: 10.1038/d41586-018-07034-4 https://www.nature.com/articles/d41586-018-07034-4

Lasica, J. D. (2003). 'Blogs and journalism need each other'. *Nieman reports*, 57(3), 70-74. 15 September 2003. https://niemanreports.org/articles/blogs-and-journalism-need-each-other/

Latvian Public Broadcasting. (2016). 'Latvia shuts down Sputnik propaganda website'. LSM.LV, 29 March 2016. https://eng.lsm.lv/article/society/society/latvia-shuts-down-sputnik-propaganda-website.a175627/

Lazer, D., Baum, M.,Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). 'Combating Fake News: An Agenda for Research and Action'. https://shorensteincenter.org/combating-fake-news-agenda-for-research/

Leathern, R. (2020). 'Expanded Transparency and More Controls for Political Ads'. *Facebook Newsroom*, 09 January 2020. https://about.fb.com/news/2020/01/political-ads/

Lee, D. (2019a). 'Key fact-checkers stop working with Facebook'. *BBC*, 02 February 2019. https://www.bbc.co.uk/news/technology-47098021

Lee, D. (2019b). 'Matter of fact-checkers: Is Facebook winning the fake news war?' *BBC*, 02 April 2019. https://www.bbc.com/news/technology-47779782

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N. de Vreese, C. H. (2019). 'Platform ad archives: promises and pitfalls'. *Internet Policy Review*, 8(4). DOI: 10.14763/2019.4.1421 https://policyreview.info/articles/analysis/platform-ad-archives-promises-and-pitfalls

LeFigaro. (2019). 'Loi contre les «fake news»: Twitter bloque une campagne du gouvernement'. *Économie*, 02 April 2019. https://www.lefigaro.fr/flash-eco/loi-contre-les-fake-news-twitter-bloque-une-campagne-du-gouvernement-20190402

Legum, J. (2019). 'Facebook says Trump can lie in his Facebook ads'. *Popular Information*, 3 October 2019. https://popular.info/p/facebook-says-trump-can-lie-in-his

Leibovich, A. (2019). 'Dark Marketing 101: The Change to How Brands Communicate Online'. *Target Marketing*, 26 February 2019. https://www.targetmarketingmag.com/article/dark-marketing-101-change-how-brands-communicate-online/

Leibowicz, C. (2019). 'Protecting Public Discourse from AI-Generated Mis/Disinformation'. Partnership on AI blog, 17 June 2019. https://www.partnershiponai.org/protecting-public-discourse-from-ai-generated-misdisinformation/

Leveson, Lord Justice. (2012). 'An Inquiry into the culture, practices and ethics of the Press'. HC, November 2012. https://www.gov.uk/government/publications/leveson-inquiry-report-into-the-culture-practices-and-ethics-of-the-press

Levien, R. & Aiken, A. (1998). 'Attack resistant trust metrics for public key certification'. In the 7th USENIX Security Symposium, San Antonio, Texas, January 1998. https://www.usenix.org/legacy/publications/library/proceedings/sec98/full_papers/levien/levien.pdf

Levin, A. (2018). ''They don't care': Facebook factchecking in disarray as journalists push to cut ties'. *The Guardian*, 13 December 2018. https://www.theguardian.com/technology/2018/dec/13/they-dont-care-facebook-fact-checking-in-disarray-as-journalists-push-to-cut-ties

Levush, R. (2019). 'Initiatives to Counter Fake News: Israel'. *Library of Congress*, April 2019. https://www.loc.gov/law/help/fake-news/israel.php

Levy, S. (2020). 'Why Mark Zuckerberg's Oversight Board May Kill His Political Ad Policy'. *Wired*, 28 January 2020. https://www.wired.com/story/facebook-oversight-board-bylaws/

Lewis, P. (2018). 'Fiction is outperforming reality': how YouTube's algorithm distorts truth'. *The Guardian*, 02 February 2018. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

Liao, Q. V., & Fu, W. T. (2013). 'Beyond the filter bubble: Interactive effects of perceived threat and topic involvement on selective exposure to information'. *In Proceedings of the SIGCHI conference on human factors in computing systems* (*CHI'13*), *2359-2368. ACM*, April 2013. https://doi.org/10.1145/2470654.2481326

Likhachev, N. (2018). 'Как работает политическая реклама во «ВКонтакте»: Собянину можно, Собчак и Навальному — нет'. *TJournal*, 05 January 2018. https://tjournal.ru/news/64619-kak-rabotaet-politicheskaya-reklama-vo-vkontakte-sobyaninu-mozhno-sobchak-i-navalnomu-net

LINE. (2020). '違反投稿への対応'.*Transparency Report*, 15 January 2020. https://linecorp.com/ja/security/moderation/2019h1

LINE. (2019a). 'お役立ち情報をお届け!「スマートチャンネル」を活用してみよう'. 01 April 2019. http://official-blog.line.me/ja/archives/78619192.html

LINE. (2019b). '【公式】LINE広告とは | サービス概要・特長まとめ'. 24 October 2019. https://www.linebiz.com/jp/column/technique/20191024/

Livingstone, S. (2020). 'Coronavirus and #fakenews: what should families do?' *LSE*, 26 March 2020. https://blogs.lse.ac.uk/medialse/2020/03/26/coronavirus-and-fakenews-what-should-families-do/

Livingstone, S., Byrne, J. & Carr, J. (2016). 'One in Three: Internet Governance and Children's Rights'. *Innocenti Discussion Papers* no. 2016-01, UNICEF Office of Research, January 2016. https://www.unicef-irc.org/publications/795-one-in-three-internet-governance-and-childrens-rights.html

Livsier, L. (2019) 'Ministry to revoke licences of media over fake news'. *Khmer Times*, August 13, 2019. https://www.khmertimeskh.com/632885/ministry-to-revoke-licences-of-media-over-fake-news/

Llansó, E., van Hoboken, J., Leerssen, P. & Harambam, J. (2020). 'Artificial Intelligence, Content Moderation, and Freedom of Expression'. (*Working Paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression*). 26 February 2020. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Loewenstein, G. (1994). 'The psychology of curiosity: A review and reinterpretation'. *Psychological Bulletin 116, 1.*

Lomas, N. (2020). 'Facebook's latest 'transparency' tool doesn't offer much — so we went digging'. *Tech Crunch*, 25 February 2020. https://techcrunch.com/2020/02/25/facebooks-latest-transparency-tool-doesnt-offer-much-so-we-went-digging/

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). ‚How behavioural sciences can promote truth, autonomy and democratic discourse online'. *Nature Human Behaviour*, 15 June 2020. https://doi.org/10.1038/s41562-020-0889-7

LSE. (2018). 'Tackling the Information Crisis: A Policy Framework for Media System Resilience'. http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

Lyons, T. (2018a). 'How is Facebook's Fact-Checking Program Working?' https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking/

Lyons, T. (2018b). 'New Research Shows Facebook Making Strides Against False News'. https://about.fb.com/news/2018/10/inside-feed-michigan-lemonde/

Lypshultz, J. H. (2018). 'Free Expression in the Age of the Internet: Social and Legal Boundaries'. *Routledge*.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). 'Hate speech detection: Challenges and solutions'. *PloS one*, 20 August 2019. 14(8). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152

MacCarthy, M. (2020). 'Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry'. One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. 12 February 2020. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Transparency_TWG_MacCarthy_Feb_2020.pdf

MacGuill, D. (2018). 'Did Facebook Flag the Declaration of Independence as Hate Speech?' *Snopes*, 06 July 2018. https://www.snopes.com/fact-check/facebook-declaration-of-independence-hate-speech/

Magdy, S. (2019). 'Egypt tightens restrictions on media, social networks'. *AP*, 19 March 2019. https://apnews.com/1540f1133267485db356db1e58db985b

Malaysian Communications and Multimedia Commission (2020). 'Press Release: Four Individuals Detained For Spreading Fake News On The Novel Coronavirus (2019-nCov) Outbreak'. *Malaysian Communications and Multimedia Commission*, 29 January 2020. https://www.mcmc.gov.my/en/media/press-releases/press-release-four-individuals-detained-for-sprea

Mantzarlis, A. (2017). 'Repetition boosts lies — but could help fact-checkers, too'. *Poynter*, 30 May 2017. https://www.poynter.org/news/repetition-boosts-lies-could-help-fact-checkers-too,

Mantzarlis, A. (2020). 'COVID-19: $6.5 million to help fight coronavirus misinformation'. *Google News Lab*, 02 April 2020. https://www.blog.google/outreach-initiatives/google-news-initiative/covid-19-65-million-help-fight-coronavirus-misinformation/

Maréchal, N. & Biddle, E. R. (2020). 'It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge'. *Open Technology Institute*. https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/

Maréchal, N., MacKinnon, R. & Biddle, E. R. (2020). 'Getting to the Source of Infodemics: It's the Business Model'. *Open Technology Institute*. https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/

Marra, F, Gragnaniello, D, Verdoliva, L. & Poggi, G. (2018). 'Do GANs leave artificial fingerprints?' 31 December 2018. https://arxiv.org/pdf/1812.11842.pdf

Marsden, C. & Meyer, T. (2019). 'Regulating disinformation with artificial intelligence: Effects of disinformation initiatives on freedom of expression and media pluralism. European Parliamentary Research Service, March 2019. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf

Martens, B., Aguiar Wicht, L., Gomez-Herrera, M. E. & Mueller-Langer, F. (2018). 'The digital transformation of news media and the rise of disinformation and fake news', *EUR - Scientific and Technical Research Reports, JRC Technical Reports, JRC Digital Economy Working Paper 2018-02*, April 2018. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/digital-transformation-news-media-and-rise-disinformation-and-fake-news#

Martin, D. A., & Shapiro, J. N. (2019). 'Trends in Online Foreign Influence Efforts'. Working Paper, Princeton University, 2019. https://scholar.princeton.edu/sites/default/files/jns/files/trends_in_foreign_influence_efforts_2019jul08_0.pdf

Masih, N., Irfan, S. & Slater, J. (2019). 'India's Internet shutdown in Kashmir is the longest ever in a democracy'. *The Washington Post*, 16 December 2019. https://www.washingtonpost.com/world/asia_pacific/indias-internet-shutdown-in-kashmir-is-now-the-longest-ever-in-a-democracy/2019/12/15/bb0693ea-1dfc-11ea-977a-15a6710ed6da_story.html

Masnick, M. (2003). 'Photo of Streisand Home Becomes An Internet Hit'. *Tech Dirt*, 24 June 2003. https://www.techdirt.com/articles/20030624/1231228.shtml

Mathur, S. (2019). 'I & B team to counter fake news'. *The Times of India*, 16 November 2019. https://timesofindia.indiatimes.com/india/ib-team-to-counter-fake-news/articleshow/72079340.cms?from=mdr&utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Mayhew, F. (2020). 'Guardian's head of investigations: 'In times of crisis you need more accountability journalism, not less'. *Press Gazette*, 27 May 2020. https://pressgazette.co.uk/guardian-head-of-investigations-paul-lewis-interview-coronavirus/

Mays, H. (2019). 'Facebook bans Tories' 'distorted' party advert with Laura Kuenssberg'. *The Guardian*, 01 December 2019. https://www.theguardian.com/politics/2019/dec/01/facebook-bans-tories-distorted-party-advert-with-laura-kuenssberg

McLaughlin, T. (2018). 'How WhatsApp Fuels Fake News and Violence in India'. *Wired*, 12 December 2018. https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/

McNamee, R. (2020). 'Social Media Platforms Claim Moderation Will Reduce Harassment, Disinformation and Conspiracies. It Won't'. *Time*, 24 June 2020. https://time.com/5855733/social-media-platforms-claim-moderation-will-reduce-harassment-disinformation-and-conspiracies-it-wont/

Melo, P. F., Messias, J., Resende, G., Garimella, V. R., Almeida, J. M., & Benevenuto, F. (2019). 'WhatsApp Monitor: A Fact-Checking System for WhatsApp'. In *Proceedings of the Int'l AAAI Conference on Weblogs and Social Media* (*ICWSM'19*). June 2019. https://people.mpi-sws.org/~johnme/pdf/melo-icwsm-2019-demo.pdf

Mexico INE (2019). 'Presenta INE modelo de combate a la desinformación en elecciones 2018-2019'. Press Release, 18 July 2019. https://centralelectoral.ine.mx/2019/07/18/presenta-ine-modelo-combate-la-desinformacion-elecciones-2018-2019/

Mezaris, V. Nixon, L. Papadopoulos, S. & Teyssou, D. (2019). 'Video Verification in the Fake News Era'. Springer.

Milano, B. (2019). 'Israeli Supreme Court Justice on combatting propaganda in elections'. *Harvard Law Today*, 29 October 2019. https://today.law.harvard.edu/israeli-supreme-court-justice-on-combatting-propaganda-in-elections/

Mitchell, A., Gottfried, J., Barthel, M. & Sumida, N. (2018). 'Distinguishing Between Factual and Opinion Statements in the News'. *Pew Research Center*, 18 June, 2018. https://www.journalism.org/wp-content/uploads/sites/8/2018/06/PJ_2018.06.18_fact-opinion_FINAL.pdf

Mitchell, A., Gottfried, J., Stocking, G. Walker, M. & Fedeli, S. (2019). 'Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed'. *Pew Research Centre*, Journalism & Media, 05 June 2019. https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Molina, K. (2016). 'Indonesian Electronic Information and Transactions Law Amended'. *White & Case*, 15 December 2016. https://www.whitecase.com/publications/alert/indonesian-electronic-information-and-transactions-law-amended

Monaco, N. & Nyst, C. (2018). 'State-Sponsored Trolling. How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns'. *IFTF*. http://www.iftf.org/statesponsoredtrolling/

Moore, N. (2020). 'Study Finds More COVID-19 Cases Among Viewers Of Fox News Host Who Downplayed The Pandemic'. *Wbez*, 30 April 2020. https://www.wbez.org/stories/study-finds-more-covid-19-cases-among-viewers-of-fox-news-host-who-downplayed-the-pandemic/83dffd03-6013-47c8-b012-0b1cc5377e18

Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., Flynn, P. J., Rocha, A. & Scheirer, W. J. (2018). 'Image Provenance Analysis at Scale'. 19 January 2018. https://arxiv.org/abs/1801.06510

Mosseri, A. (2016). 'Addressing Hoaxes and Fake News'. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

MSB. (2020). 'Countering information influence activities. A handbook for communicators'. Swedish Civil Contingencies Agency. https://rib.msb.se/filer/pdf/28698.pdf

Mullen, A. (2016). 'Leave vs Remain: the Digital Battle' In 'EU Referendum Analysis 2016: Media, Voters and the Campaign'. Jackson, D, Thorsen, E, Wring, D. Loughborough University Center for the Study of Journalism, Culture and Community. http://www.meandeurope.co.uk/wp-content/uploads/EU-Referendum-Analysis-2016-Jackson-Thorsen-and-Wring-v2.pdf

Munson, S. A., Lee, S. Y. & Resnick, P. (2013). 'Encouraging Reading of Diverse Political Viewpoints with a Browser Widget'. In International conference on weblogs and social media (ICWSM). Boston. https://webcache.googleusercontent.com/search?q=cache:TcCE3Axx_9MJ:https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6119/6381+&cd=1&hl=en&ct=clnk&gl=uk&client=firefox-b-e

Myanmar Penal Code. (1861). https://www.burmalibrary.org/docs6/MYANMAR_PENAL_CODE-corr.1.pdf

Myanmar Telecommunications Law. (2013). 'The Telecommunications Law (The Pyidaungsu Hluttaw Law No. 31, 2013) The 4th Waxing Day of Thadingyut, 1375 M.E., 8 October 2013. https://www.burmalibrary.org/docs23/2013-10-08-Telecommunications_Law-en.pdf

Nagulendra, S., & Vassileva, J. (2014). 'Understanding and controlling the filter bubble through interactive visualization: a user study'. *In Proceedings of the 25th ACM conference on Hypertext and social media* (*pp. 107-115*), September 2014. https://www.researchgate.net/publication/266660926_Understanding_and_controlling_the_filter_bubble_through_interactive_visualization_A_user_study

National Literacy Trust. (2018). 'Fake news and critical literacy: final report'. 11 June 2018. https://literacytrust.org.uk/research-services/research-reports/fake-news-and-critical-literacy-final-report/

NATO Stratcom COE (2019). '*About Us*', https://www.stratcomcoe.org/about-us

Neate, R. (2017). 'Bell Pottinger faces hearing over claims it stirred racial tension in South Africa'. *The Guardian*, 13 August 2017. https://www.theguardian.com/world/2017/aug/13/bell-pottinger-pr-industry-hearing-secret-south-africa-campaign

Nechushtai, E., & Lewis, S.C. (2018). 'What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations'. *Computers in Human Behavior*, 90, 298-307. https://doi.org/10.1016/j.chb.2018.07.043

NED. (2018). Comparative Responses to the Global Disinformation Challenge. National Endowment for Democracy. October 4-5, 2018

Nelson, R.A. (1996). 'A Chronology and Glossary of Propaganda in the United States, Westport, Connecticut', p232-233. *Greenwood Press*, ISBN 0313292612

Netherlands Policy Report for Online Disinformation. (2018). *Dropbox*, July, 2018. https://www.dropbox.com/s/99iza9kmbwjbels/20180718_rapport_onlinedesinformatieNL.pdf?dl=0

Newman, L. H. (2019a). 'To Fight Deepfakes, Researchers Built a Smarter Camera'. *Wired*, 28 May 2019. https://www.wired.com/story/detect-deepfakes-camera-watermark/

Newman, N. (2019b). 'Executive Summary and Key Findings of the 2019 Report'. *Reuters, University of Oxford, Digital News Report*. http://www.digitalnewsreport.org/survey/2019/overview-key-findings-2019/

Newman, N. & Fletcher, R. (2017). 'Bias, Bullshit and Lies: Audience Perspectives on Low Trust in the Media'. *Reuters*. https://reutersinstitute.politics.ox.ac.uk/our-research/bias-bullshit-and-lies-audience-perspectives-low-trust-media

Newsguard. (2020a). 'Coronavirus Misinformation Tracking Center'. 20 April 2020. https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/

Newsguard. (2020b). 'NewsGuard Partners with DCMS and BT to Help Counter Spread of COVID-19 Fake News as Misinformation Peaks'. 27 March 2020. https://www.newsguardtech.com/press/newsguard-partners-with-dcms-and-bt-to-help-counter-spread-of-covid-19-fake-news-as-misinformation-peaks/

Newton, C. (2019a). 'Why a top content moderation company quit the business instead of fixing its problems'. *The Verge*, 01 November 2019. https://www.theverge.com/interface/2019/11/1/20941952/cognizant-content-moderation-restructuring-facebook-twitter-google

Newton, C. (2019b). 'The Trauma Floor. The secret lives of Facebook moderators in America'. *The Verge*, 25 February 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2020). 'Snap will stop promoting Trump's account after concluding his tweets incited violence'. *The Verge*, 03 June 2020. https://www.theverge.com/2020/6/3/21279280/snapchat-snap-remove-trump-account-discover-promotion-incite-violence-twitter

New York Times. (2019). 'Read the Letter Facebook Employees Sent to Mark Zuckerberg About Political Ads'. 28 October 2019. https://www.nytimes.com/2019/10/28/technology/facebook-mark-zuckerberg-letter.html

New York Times. (2020). 'How China Is Reshaping the Coronavirus Narrative'. *You Tube*, 18 March 2020. https://www.youtube.com/watch?v=LwqhvRcBrK4&feature=youtu.be

New Zealand Parliament Justice Committee. (2019). Inquiry into the 2017 General Election and 2016 Local Elections. https://www.parliament.nz/resource/en-NZ/SCR_93429/5dd1d57eeba54f36bf9f4da96dba12c073ed7ad8

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*. 2:175. https://doi.org/10.1037/1089-2680.2.2.175

Nigerian Senate. (2019). Nigeria Protection from Internet Falsehood and Manipulation Bill. (2019). https://ndlink.org/wp-content/uploads/2019/11/Protection-from-Internet-Falsehood-and-Manipulation-Bill-2019.pdf

Nimmo, B. (2019). 'UK Trade Leaks and Secondary Infektion: New Findings and Insights from a Known Russian Operation'. *Graphika*, December 2019. https://graphika.com/uploads/Graphika%20Report%20-%20UK%20Trade%20Leaks%20&%20Secondary%20Infektion.pdf

Nimmo, B., Shawn Eib, C. & Tamora, L. (2019a). 'Cross-Platform Spam Network Targeted Hong Kong Protests. "Spamouflage Dragon" used hijacked and fake accounts to amplify video content'. *Graphika*, September 2019. https://swank-range.cloudvent.net/uploads/Graphika%20Report%20-%20Cross-Platform%20Spam%20Network%20Targeted%20Hong%20Kong%20Protests.pdf

Nimmo, B., Shawn Eib, C., Tamora, L., Johnson, K., Smith, I. (from Graphika), Buziashvili, E., Kann, A., Karan, K., Ponce de Leon Rosas, E. & Rizzuto, M. (from DFRLab) (2019b). '#OperationFFS: Fake Face Swarm'. *Graphika* & *DFRLab* 20 December 2019. https://graphika.com/reports/operationffs-fake-face-swarm/

Norris, P., Cameron, S., & Wynter, T. (2019). Electoral Integrity in America. Securing Democracy. Oxford University Press.

Nuñez, M. (2020). 'Snap Stock Drops 11% After Revenue Falls Short Of Expectations'. *Forbes*, 04 February 2020. https://www.forbes.com/sites/mnunez/2020/02/04/snap-stock-drops-14-after-revenue-falls-short-of-expectations/#5e7c1bf0448c

Nygren, T. & Guath, M. (2019). 'Swedish teenagers' difficulties and abilities to determine digital news credibility'. *Sciendo*, 19 February 2019. https://doi.org/10.2478/nor-2019-0002

Nygren, T. & Guath, M. (2020). 'A Multidisciplinary Look at Knowledge Resistance'. *YouCheck*, 14 January 2020. http://project-youcheck.com/a-multidisciplinary-look-at-knowledge-resistance/

Nyhan, B. (2012). Columbia Journalism Review: Does Fact-Checking Work? False Statements are Wrong Metric. https://archives.cjr.org/united_states_project/does_fact-checking_work_false.php

Nyhan, B. & Reifler, J. (2006). 'When Corrections Fail: The persistence of political misperceptions' Forthcoming, *Political Behavior*. https://www.dartmouth.edu/~nyhan/nyhan-reifler.pdf.

Nyhan, B & Reifler, J. (2010). 'When Corrections Fail: The Persistence of Political Misperceptions'. *Political Behavior*, 30 March 2010. https://link.springer.com/article/10.1007%2Fs11109-010-9112-2

Nyhan, B.& Reifler, J. (2012). 'Misinformation and Fact-checking: Research Findings from Social Science'. http://web.archive.org/web/20150226014434/http://mediapolicy.newamerica.net/sites/newamerica.net/files/policydocs/Misinformation_and_Fact-checking.pdf

OAS. (2017). 'Office Of The Special Rapporteur For Freedom Of Expression Expresses Serious Concern Over The Enactment Of The "Anti-Hate Law" In Venezuela And Its Effects On Freedom Of Expression And Freedom Of The Press'. Press Release, 10 November 2017. http://www.oas.org/en/iachr/expression/showarticle.asp?artID=1082&lID=1

OAS. (2019). IACHR (Inter-American Commission on Human Rights) & RFOE (Special Rapporteur for Freedom of Expression). Guide to guarantee freedom of expression regarding deliberate disinformation in electoral contexts. http://www.oas.org/en/iachr/expression/publications/Guia_Desinformacion_VF%20ENG.pdf

O'Brien, C. & Kelly, F. (2018). 'Google Bans Online Ads on Abortion Referendum'. *The Irish Times*, 9 May 2018. https://www.irishtimes.com/business/media-and-marketing/google-bans-online-ads-on-abortion-referendum-1.3489046

Ofcom. (2018a). 'News consumption in the UK'. 25 July 2018. https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption.

Ofcom. (2018b). 'Addressing harmful online content'. 18 September 2018. https://www.ofcom.org.uk/phones-telecoms-and-internet/information-for-industry/online-policy-research/addressing-harmful-online-content

Oliphant, R. (2017). 'Ukraine bans Russian social networks in sweeping expansion of sanctions'. *The Telegraph*, 16 May 2017. https://www.telegraph.co.uk/news/2017/05/16/ukraine-bans-russian-social-networks-sweeping-expansion-sanctions/

Ong, J. C. & Cabañes, J. V. A. (2018). 'Architects of Networked Disinformation. Behind the Scenes of Troll Accounts and Fake News Production in the Philippines'. *The Newton Tech4Dev Network*, 17 February 2018. https://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf

Ong, J. C. & Cabañes, J. V. A. (2019). 'Politics and Profit in the Fake News Factory: Four Work Models of Political Trolling in the Philippines'. *NATO Strategic Communications Centre of Excellence*, November 2019. https://stratcomcoe.org/four-work-models-political-trolling-philippines.

OSCE. (2017). UN/OSCE/OAS/ACHPR Rapporteurs on Freedom of Expression. (2017). *Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda*'. https://www.osce.org/fom/302796?download=true

Osmundsen, M., Bor, A., Vahlstrup, P.B., Bechmann, A. & Petersen, M.B. (2020). 'Partisan Polarization is the Primary Psychological Motivation Behind "Fake News" Sharing on Twitter'. 25 March 2020, https://psyarxiv.com/v45bk/

Pakistani Prevention of Electronic Crimes Act of 2016. http://www.na.gov.pk/uploads/documents/1470910659_707.pdf

Pamment, J., Nothaft, H., Agardh-Twetman, H. & Fjällhed. (2018). 'Countering Information Influence Activities: A handbook for communicators'. https://www.msb.se/RibData/Filer/pdf/28698.pdf

Paquette, D. (2019). 'Nigeria's 'fake news' bill could jail people for lying on social media. Critics call it censorship.' *The Washington Post*, *25* November 2019. https://www.washingtonpost.com/world/africa/nigerias-fake-news-bill-could-jail-people-for-lying-on-social-media-critics-call-it-censorship/2019/11/25/ccf33c54-0f81-11ea-a533-90a7becf7713_story.html

Palma, S., Munshi, N. & Reed, J. (2020). 'Singapore 'falsehoods' law shows perils of fake news fight'. *Financial Times*, 04 February 2020. https://www.ft.com/content/e50eb042-3db3-11ea-a01a-bae547046735

Parkinson, H. J. (2016). 'Click and Elect: How Fake News Helped Donald Trump Win a Real Election'. *The Guardian*, 14 November 2016. https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies

Partnership on AI. (2020). 'The Deepfake Detection Challenge: Insights and Recommendations for AI and Media Integrity'. 12 March 2020. https://www.partnershiponai.org/a-report-on-the-deepfake-detection-challenge/

Pasternack, A. (2020). 'Facebook is quietly pressuring its independent fact-checkers to change their rulings'. *Fast Company*, 20 August 2020. https://www.fastcompany.com/90538655/facebook-is-quietly-pressuring-its-independent-fact-checkers-to-change-their-rulings

Paul, R. (2018). 'Bangladesh shuts down high-speed mobile internet on election eve'. *Reuters*, 29 December 2018. https://www.reuters.com/article/us-bangladesh-election-internet/bangladesh-shuts-down-high-speed-mobile-internet-on-election-eve-idUSKCN1OS0AR

Penney, V. (2020). 'How Facebook Handles Climate Disinformation'. *New York Times*, 14 July 2020. https://www.nytimes.com/2020/07/14/climate/climate-facebook-fact-checking.html

Pennycook, G., Bear, A. Collins, E. & Rand, D.G. (2020) "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings." Working paper, Last revised: 16 Jan 2020. http://dx.doi.org/10.2139/ssrn.3035384

Pennycook, G. & Rand, D.G. (2019). 'Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning.' *Cognition* 188, July 2019. pp39-50. https://doi.org/10.1016/j.cognition.2018.06.011

People's Daily. (2013). '国家互联网信息办部署打击网络谣言'. 03 May 2013. http://politics.people.com.cn/n/2013/0503/c1001-21348755.html

Perraudin, F. (2019). 'Twitter accuses Tories of misleading public with 'factcheck' foray'. *The Guardian*, 20 November 2019. https://www.theguardian.com/politics/2019/nov/20/twitter-accuses-tories-of-misleading-public-in-factcheck-row

Philippines Revised Penal Code, Act no 10951. (2017). https://www.officialgazette.gov.ph/downloads/2017/08aug/20170829-RA-10951-RRD.pdf

Philippines Senate. (2019). Bill no 9, Anti-False Content Bill. An Act prohibiting the publication and proliferation of false content on the Philippine internet, providing measures to counteract its effects and prescribing penalties therefore. https://senate.gov.ph/lisdata/3022527054!.pdf

Pichai, S. (2020). 'COVID-19: How we're continuing to help'. *Google*, 15 March 2020. https://blog.google/inside-google/company-announcements/covid-19-how-were-continuing-to-help/

PKULaw. (2017). 'Criminal Law of the People's Republic of China (2017 Amendment PKULAW Version) [Effective]'. 04 November 2017. http://en.pkulaw.cn/display.aspx?cgid=703dba7964330b85bdfb&lib=law

Plasilova, I., Hill, J., Carlberg, M., Goubet, M. & Procee, R. (2020). 'STUDY FOR THE "Assessment of the implementation of the Code of Practice on Disinformation". European Commission, May 2020. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=66649

PolitiFact (2018). 'Here's our statement on a recent story in The Guardian about third-party fact-checking on Facebook. Four other fact-checking groups signed as well'. *Twitter*, 13 December 2018. https://twitter.com/politifact/status/1073309698894585856?s=11

Pollock, S. (2019). 'Facebook fined €2m for under-reporting its complaints figures'. *The Times*, 07 July 2019. https://www.thetimes.co.uk/article/facebook-fined-2m-for-under-reporting-its-complaints-figures-7fhg26dct

Polonski, V. (2016). 'Impact of social media on the outcome of the EU referendum'. In EU Referendum Analysis 2016: Media, Voters and the Campaign. Jackson, D, Thorsen, E, Wring, D. Loughborough University Center for the Study of Journalism, Culture and Community. http://www.referendumanalysis.eu/

Pomares, J. & Guzman, N. (2015). 'The hardest check: Measuring the impact of fact-checking'. https://www.poynter.org/wp-content/uploads/2015/10/The-hardest-check-1.pdf

Posetti, J. (2013). 'The 'Twitterisation' of investigative journalism' in S. Tanner & N. Richardson (Eds.), Journalism Research and Investigation in a Digital World (pp. 88-100): *Oxford University Press*, Melbourne. https://ro.uow.edu.au/cgi/viewcontent.cgi?article=2765&context=lhapapers

Posetti, J. (2017). 'Fighting Back Against Prolific Online Harassment: Maria Ressa'. In '*An Attack on One is an Attack on All*. Successful Initiatives To Protect Journalists and Combat Impunity', Larry Kilman, UNESCO. https://en.unesco.org/sites/default/files/an_attack_on_on_is_an_attack_on_all_chapter_8.pdf

Posetti, J. (2017b). 'Protecting Journalism Sources in the Digital Age'. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000248054

Posetti, J. (2018a). 'Combatting online abuse: when journalists and their sources are targeted'. https://en.unesco.org/sites/default/files/module_7.pdf

Posetti, J. (2018b). ''News industry transformation: digital technology, social platforms and the spread of misinformation and disinformation' in Ireton & Posetti (eds) Journalism, 'Fake News' & Disinformation. (UNESCO) https://en.unesco.org/fightfakenews

Posetti, J. (2018c). 'Time to step away from the 'bright, shiny things'? Towards a sustainable model of journalism innovation in an era of perpetual change'. *Journalism Innovation Project, Reuters & the University of Oxford*. November 2018. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-11/Posetti_Towards_a_Sustainable_model_of_Journalism_FINAL.pdf

Posetti, J. (2020). 'Journalists like Maria Ressa face death threats and jail for doing their jobs. Facebook must take its share of the blame'. *CNN*, 02 July 2020. https://edition.cnn.com/2020/06/30/opinions/maria-ressa-facebook-intl-hnk/index.html

Posetti, J. & Bontcheva, K. (2020a). 'Disinfodemic – Deciphering COVID-19 disinformation'. Policy Brief 1. *UNESCO*, April 2020. https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf

Posetti, J. & Bontcheva, K. (2020b). 'Disinfodemic – Dissecting responses to COVID-19 disinformation'. Policy Brief 2. *UNESCO*, April 2020. https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf

Posetti, J. & Matthews, A. (2018). 'A Short Guide to the History of 'Fake News' and Disinformation. A Learning Module for Journalists and Journalism Educators'. *International Center For Journalists*, 23 July 2018. https://www.icfj.org/sites/default/files/2018-07/A%20Short%20Guide%20to%20History%20of%20Fake%20News%20and%20Disinformation_ICFJ%20Final.pdf

Posetti, J., Simon, F. & Shabbir, N. (2019a). 'Lessons in Innovation: How International News Organisations Combat Disinformation through Mission-Driven Journalism'. *Reuters*, April 2019. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-04/Posetti_Lessons_in_Innovation_FINAL.pdf

Posetti, J., Simon, F. & Shabbir, N. (2019b). 'What if Scale Breaks Community? Rebooting Audience Engagement When Journalism is Under Fire'. *Reuters*, October 2019. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-10/Posetti%20What%20if%20FINAL.pdf

Powell, A. (2020). 'Washington Post's Baron sends along the Class of 2020 with message that facts, truth matter'. *The Harvard Gazette*, 28 May 2020. https://news.harvard.edu/gazette/story/2020/05/martin-barons-message-to-class-of-2020-facts-and-truth-matter/?utm_medium=socialmedia&utm_source=hkstwitter

Power, J. (2020). 'Coronavirus vaccine: anti-vax movement threatens Asian recovery'. *SCMP, The Coronvirus Pandemic*, 01 June 2020. https://www.scmp.com/week-asia/health-environment/article/3086761/coronavirus-vaccine-anti-vax-movement-threatens-asian

Poynter. 'Commit to transparency — sign up for the International Fact-Checking Network's code of principles'. *IFCN, Poynter*. https://ifcncodeofprinciples.poynter.org/

Prager, A. (2019). 'Germany wants to fine Facebook over hate speech reporting'. *EurActiv*, 03 July 2019. https://www.euractiv.com/section/digital/news/germany-wants-to-fine-facebook-over-online-hate-speech-reporting/

Privacy International. (2020). 'No, Facebook is not telling you everything'. 24 February 2020. https://privacyinternational.org/long-read/3372/no-facebook-not-telling-you-everything

Proctor, K. (2020). 'UK anti-fake news unit dealing with up to 10 false coronavirus articles a day'. 30 March 2020. https://www.theguardian.com/world/2020/mar/30/uk-anti-fake-news-unit-coronavirus

Public Media Alliance. (2020). 'Emergency funds for US public broadcasting'. 01 April 2020. https://www.publicmediaalliance.org/emergency-funds-for-us-public-broadcasting/

Quadir, S. (2018). 'Tough New Bangladesh Measure Becomes Law, Seen Curbing Free Speech', *Reuters*, 28 October 2018. https://www.reuters.com/article/us-bangladesh-politics-journalism/tough-new-bangladesh-measure-becomes-law-seen-curbing-free-speech-idUSKCN1MI0NE

Quattrociocchi, W., Scala, A. & Sunstein, C. R. (2016). 'Echo Chambers on Facebook'. *Social Science Research Network*, 15 June 2016. https://papers.ssrn.com/abstract=2795110

Qui, L. (2020). 'Analyzing the Patterns in Trump's Falsehoods About Coronavirus'. *The New York Times*, 27 March 2020. https://www.nytimes.com/2020/03/27/us/politics/trump-coronavirus-factcheck.html

Qui, S. & Woo, R. (2018). 'China launches platform to stamp out 'online rumors'. *Reuters*, 30 August 2018. https://www.reuters.com/article/us-china-internet/china-launches-platform-to-stamp-out-online-rumors-idUSKCN1LF0HL

Quinn, C. (2020). 'Hungary's Orban Given Power to Rule By Decree With No End Date'. *Foreign Policy*, 31 March 2020. https://foreignpolicy.com/2020/03/31/hungarys-orban-given-power-to-rule-by-decree-with-no-end-date/

Rahman, S., Tully, P. & Foster, L. (2019). 'Attention is All They Need: Combatting Social Media Information Operations With Neural Language Models'. *Fire Eye*, 14 November 2019. https://www.fireeye.com/blog/threat-research/2019/11/combatting-social-media-information-operations-neural-language-models.html

Ranking Digital Rights. (2019). '2019 RDR Corporate Accountability Index'. https://rankingdigitalrights.org/index2019/

Ranking Digital Rights. (2020). 'RDR Corporate Accountability Index: Transparency and accountability standards for targeted advertising and algorithmic systems: Pilot Study and Lessons Learned', 16 March 2020. https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf

Rappler Research Team. (2018). 'Tip of the iceberg: Tracing the network of spammy pages in Facebook takedown'. 27 October 2018. https://www.rappler.com/newsbreak/investigative/215256-tracing-spammy-pages-network-facebook-takedown

Rauchfleisch, A. & Kaiser, J. (2020). 'The False Positive Problem of Automatic Bot Detection in Social Science Research'. *Berkman Klein Center Research Publication No. 2020-3*, March 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3565233

Read, M. (2016). 'Donald Trump won because of Facebook'. *New York Magazine*, 09 November 2016. https://nymag.com/intelligencer/2016/11/donald-trump-won-because-of-facebook.html

Reda, J. (2017). 'When filters fail: These cases show we can't trust algorithms to clean up the internet'. 28 September 2017. https://juliareda.eu/2017/09/when-filters-fail/

Redação Exame. (2019). '*Após derrubar veto de Bolsonaro, Congresso instaura CPI das Fake News, EXAME*'. 2 September 2019 https://exame.abril.com.br/brasil/apos-derrubar-veto-de-bolsonaro-congresso-instaura-cpi-das-fake-news/

Reid, A. & Dotto. C. (2019). 'Thousands of misleading Conservative ads side-step scrutiny thanks to Facebook policy'. *First Draft*, 06 December 2019. https://firstdraftnews.org/latest/thousands-of-misleading-conservative-ads-side-step-scrutiny-thanks-to-facebook-policy/

Reppell, L. & Shein, E. (2019). 'Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions'. *IFES*, 26 April 2019. https://www.ifes.org/publications/disinformation-campaigns-and-hate-speech-exploring-relationship-and-programming

Repnikova, M. (2018). 'China's Lessons for Fighting Fake News'. *FP*, 06 September 2018. https://foreignpolicy.com/2018/09/06/chinas-lessons-for-fighting-fake-news/

Resnick, P. Ovadya, A. & Gilchrist, G. (2019). 'Iffy Quotient: A Platform Health Metric for Misinformation'. *Social Media Responsibility, University of Michigan*, 23 July 2019. https://csmr.umich.edu/wp-content/uploads/2019/07/UMSI-CSMR-Iffy-Quotient-Whitepaper-v2.pdf

Ressa, M. A. (2016). 'Propaganda war: Weaponizing the internet'. *Rappler*, 7 February 2019. https://www.rappler.com/nation/148007-propaganda-war-weaponizing-internet

Ressa, M. (2019). 'When journalists are under attack, democracy is under attack'. *Daily Maverick*, 30 September 2019. https://www.dailymaverick.co.za/article/2019-09-30-when-journalists-are-under-attack-democracy-is-under-attack/

Revelli, A. & Foster, L. (2020). '"Distinguished Impersonator" Information Operation That Previously Impersonated U.S. Politicians and Journalists on Social Media Leverages Fabricated U.S. Liberal Personas to Promote Iranian Interests'. *FireEye*, 12 February 2020. https://www.fireeye.com/blog/threat-research/2020/02/information-operations-fabricated-personas-to-promote-iranian-interests.html

RFE/RL. (2018a). 'Belarus Passes Legislation Against 'Fake News' Media'. 14 June 2018. https://www.rferl.org/a/belarus-assembly-passes-controversial-fake-news-media-legislation/29291033.html

RFE/RL. (2018b). 'Kazakhstan Shuts Down Independent News Site'. 28 May 2018. https://www.rferl.org/a/kazakhstan-shuts-down-independent-news-site-ratel/29254964.html

Richter, A. (2019) 'Disinformation in the media under Russian law'. *IRIS Extra, European Audiovisual Observatory*, June 2019. https://rm.coe.int/disinformation-in-the-media-under-russian-law/1680967369

Robinson, O., Coleman, A., Sardarizadeh, S. (2019). 'A Report of Anti-Disinformation Initiatives'. *OXTEC*, August 2019. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/A-Report-of-Anti-Disinformation-Initiatives

Rodriguez-Ferrand, G. (2019). 'Initiatives to Counter Fake News: Argentina'. *Library of Congress*, April 2019. https://www.loc.gov/law/help/fake-news/argentina.php#_ftnref15

Romm, T., Dwoskin, E. & Timberg, C. (2019). 'Sri Lanka's social media shutdown illustrates global discontent with Silicon Valley'. *The Washington Post*, 22 April 2019. https://www.washingtonpost.com/technology/2019/04/22/sri-lankas-social-media-shutdown-illustrates-global-discontent-with-silicon-valley/

Romm, T. Stanley-Becker, I. & Timberg, C. (2020). '"Facebook won't limit political ad targetin or stop false claims under new ad rules'. *The Washington Post*, 09 January 2020. https://www.washingtonpost.com/technology/2020/01/09/facebook-wont-limit-political-ad-targeting-or-stop-pols-lying/

Rosen, G. (2020). 'Investments to Fight Polarization'. *Facebook Newsroom*, 27 May 2020. https://about.fb.com/news/2020/05/investments-to-fight-polarization/

Rosen, G. & Lyons, T. (2019). 'Remove, Reduce, Inform: New Steps to Manage Problematic Content'. *Facebook Newsroom*, 10 April 2019. https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/

Rosling, H. (2018). 'Factfulness: Ten Reasons We're Wrong About The World–And Why Things Are Better Than You Think'. Sceptre.

Rössler, A., Cozzolino, D., Veroliva, L., Riess, C. Thies, J. & Neiβner, M. (2018). 'FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces'. 24 March 2018. https://arxiv.org/abs/1803.09179

Roth, Y. (2019). 'Information operations on Twitter: principles, process, and disclosure'. *Twitter Blog*, 13 June 2019. https://blog.twitter.com/en_us/topics/company/2019/information-ops-on-twitter.html

Roth, Y. & Achuthan, A. (2020). 'Building rules in public: Our approach to synthetic & manipulated media'. *Twitter Blog*, 04 February 2020. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

Roucaute, D. (2017). '« Le Monde » s'engage dans l'éducation à l'information'. *Le Monde*, 02 February, 2017 (updated 27 November 2017). https://www.lemonde.fr/les-decodeurs/article/2017/02/02/le-monde-s-engage-dans-l-education-a-l-information_5073215_4355770.html

RSF. (2018). 'Online Harassment of Journalists - Attack of the trolls'. https://rsf.org/sites/default/files/rsf_report_on_online_harassment.pdf

RSF (2019). 'Burkina Faso : l'amendement du code pénal doit être déclaré inconstitutionnel'. July 17, 2019. https://rsf.org/fr/actualites/burkina-faso-lamendement-du-code-penal-doit-etre-declare-inconstitutionnel-2

Ruan, L., Knockel, J., Q. Ng, J., & Crete-Nishihata, M. (2016). 'One App, Two Systems: How WeChat uses one censorship policy in China and another internationally'. *The Citizen Lab*, 30 November 2016. https://citizenlab.ca/2016/11/wechat-china-censorship-one-app-two-systems/

Sabbagh, D. (2020). 'Cross-Whitehall unit set up to counter false coronavirus claims'. *The Guardian*, 09 March 2020. https://www.theguardian.com/world/2020/mar/09/cross-whitehall-unit-coronavirus-disinformation

Sadek, G. (2019). 'Initiatives to Counter Fake News: Egypt'. *Library of Congress*, April 2019. https://www.loc.gov/law/help/fake-news/egypt.php#_ftn11

Sally Chan, M., Jones, C. R., Hall Jamieson, K. & Albarracin, D. (2017). 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation'. *Psychological Science*, 12 September 2017. https://journals.sagepub.com/doi/full/10.1177/0956797617714579

Sanders, E. (2020). 'Washington State Sues Facebook for "Repeatedly" Violating Campaign Finance Law'. *The Stranger*, 14 April 2020. https://www.thestranger.com/slog/2020/04/14/43407935/washington-state-sues-facebook-for-repeatedly-violating-campaign-finance-law

Sang-Hun, C. (2018). 'South Korea Declares War on 'Fake News', Worrying Government Critics'. *The New York Times*. 02 October 2018. https://www.nytimes.com/2018/10/02/world/asia/south-korea-fake-news.html

Sankin, A. (2020). 'Want to Find a Misinformed Public? Facebook's Already Done It'. *The Markup*, 23 April 2020. https://themarkup.org/coronavirus/2020/04/23/want-to-find-a-misinformed-public-facebooks-already-done-it

Satariano, A. (2018). 'Ireland's Abortion Referendum Becomes a Test for Facebook and Google', *New York Times*, 25 May 2018. https://www.nytimes.com/2018/05/25/technology/ireland-abortion-vote-facebook-google.html

Satariano, A. & Tsang, A. (2019). 'Who's Spreading Disinformation in U.K. Election? You Might Be Surprised'. *New York Times*, 10 December 2019. https://www.nytimes.com/2019/12/10/world/europe/elections-disinformation-social-media.html

Schiffrin, A. & Goodman, E. (2019). 'AI Startups and the Fight Against Online Disinformation'. *The German Marshall Fund of the United States*, 05 September 2019. http://www.gmfus.org/publications/ai-startups-and-fight-against-online-disinformation

Schmerling, R. H. (2020). 'Be careful where you get your news about coronavirus'. *Harvard Health Publishing*, 01 February 2020. https://www.health.harvard.edu/blog/be-careful-where-you-get-your-news-about-coronavirus-2020020118801

Schmitt-Beck, R. (2008). 'Bandwagon effect'. In W. Donsbach (Ed.) 'The international encyclopedia of communication', Vol. 2 (pp. 308–310). Oxford, UK. Wiley-Blackwell

Schroepfer, M. (2019). 'Creating a data set and a challenge for deepfakes'. *Facebook AI*, 05 September 2019. https://ai.facebook.com/blog/deepfake-detection-challenge/

Schulman, S. (2019). 'Development, digitalization and disinformation in Myanmar'. *U brief*. May 2019. https://www.ui.se/globalassets/ui.se-eng/publications/ui-publications/2019/ui-brief-no.-5-2019.pdf

Schulten, K. & Brown, A. C. (2017). 'Evaluating Sources in a 'Post-Truth' World: Ideas for Teaching and Learning About Fake News'. New York Times, 19 January 2017. https://www.nytimes.com/2017/01/19/learning/lesson-plans/evaluating-sources-in-a-post-truth-world-ideas-for-teaching-and-learning-about-fake-news.html

Schumaker, E. (2019). 'Why Conspiracy Theories Work so Well on Facebook'. OneZero, 05 March 2019. https://onezero.medium.com/why-conspiracy-theories-work-so-well-on-facebook-466a42af6b76

Scott, C. (2018). 'With an interactive game, the BBC is helping young people better understand the disinformation ecosystem'. Journalism.co.uk, 01 May 2018. https://www.journalism.co.uk/news/with-an-interactive-game-the-bbc-is-helping-young-people-better-understand-the-disinformation-ecosystem-/s2/a721168/

See Kit, T. (2019). 'Law to curb deliberate online falsehoods takes effect'. *CNA*. 02 October 2019. https://www.channelnewsasia.com/news/singapore/law-to-curb-deliberate-online-falsehoods-takes-effect-11962068

Sen, A. & Zadrozny, B. (2020). 'QAnon groups have millions of members on Facebook, documents show'. *NBC News*, 10 August 2020. https://www.nbcnews.com/tech/tech-news/qanon-groups-have-millions-members-facebook-documents-show-n1236317

Sénécat, A. (2018). 'Les fausses informations circulent de moins en moins sur Facebook'. *Le Monde*, 17 October 2018. https://www.lemonde.fr/les-decodeurs/article/2018/10/17/les-fausses-informations-perdent-du-terrain-sur-facebook_5370461_4355770.html

Shin, J. & Thorsen, K. (2017). 'Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media'. *Journal of Communication*, 28 February 2017. https://academic.oup.com/joc/article/67/2/233/4082394

Shu, C. & Shieber, J. (2020). 'Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation'. *TechCrunch*, 17 March 2020. https://techcrunch.com/2020/03/16/facebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation/

Silva, M., Oliveira, L. S., Andreou, A., Vaz de Melo, P. O., Goga, O. & Benevenuto, F. (2020). 'Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook'. In *Proceedings of the Web Conference* (*WWW'20*), April 2020. https://doi.org/10.1145/3366423.3380109

Silverman, C. et al. (2014). 'The Verification Handbook (European Journalism Centre). https://verificationhandbook.com/book/index.php

Silverman, C. (2016). 'Fake News Expert On How False Stories Spread And Why People Believe Them'. Fresh Air – transcript on *NPR*, 14 December 2016. https://www.npr.org/2016/12/14/505547295/fake-news-expert-on-how-false-stories-spread-and-why-people-believe-them?t=1589899554744

Silverman, C. (2017a). '5 ways scammers exploit Facebook to feed you false information'. *BuzzFeed*, 28 April 2017. https://www.buzzfeed.com/craigsilverman/how-facebook-is-getting-played

Silverman, C. (2017b). 'Facebook says its Fact-Checking Program helps reduce the spread of a fake story by 80%'. *Buzzfeed*, 11 October 2017. https://www.buzzfeednews.com/article/craigsilverman/facebook-just-shared-the-first-data-about-how-effective-its

Silverman, C. (2019). 'How A Massive Facebook Scam Siphoned Millions Of Dollars From Unsuspecting Boomers'. *Buzzfeed*, 16 October 2019. https://www.buzzfeednews.com/article/craigsilverman/facebook-subscription-trap-free-trial-scam-ads-inc

Silverman, C. (Ed.) (2020). 'Verification Handbook. For Disinformation And Media Manipulation'. *European Journalism Centre*. https://datajournalism.com/read/handbook/verification-3

Silverman, C. & Alexander, L. (2016). 'How Teens In The Balkans Are Duping Trump Supporters With Fake News'. *BuzzFeed*, 3 November 2016. https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo

Silverman, C., Lytvynenko, J. & Kung, W. (2020). 'Disinformation For Hire: How a New Breed of PR Firms Is Selling Lies Online'. *BuzzFeed*, 6 January 2020. https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms

Silverman, C., Lytvynenko, J. & Pham, S. (2017). 'These Are 50 Of The Biggest Fake News Hits On Facebook In 2017'. *Buzzfeed*, 28 December 2017. https://www.buzzfeednews.com/article/craigsilverman/these-are-50-of-the-biggest-fake-news-hits-on-facebook-in

Silverman C, Mac R, Dixit P (2020) "Facebook Is Turning A Blind Eye To Global Political Manipulation, According To This Explosive Secret Memo", Buzzfeed. September 14th, 2020 https://www.buzzfeednews.com/amphtml/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo?__twitter_impression=true

Silverman, C. & Pham, S. (2018). '*These Are 50 Of The Biggest Fake News Hits On Facebook In 2018*'. *Buzzfeed*, 28 December 2018. https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-news-hits-2018

Simon, J. (2020). 'COVID-19 is spawning a global press-freedom crackdown'. *CJR*, 25 March 2020. https://www.cjr.org/analysis/coronavirus-press-freedom-crackdown.php

Singapore Democratic Party (2020). 'SDP Files Case Against MOM In High Court To Fight For What Little Democratic Space We Have Left In S'pore'. 08 January 2020. https://yoursdp.org/news/sdp-files-case-against-mom-in-high-court-to-fight-for-what-little-democratic-space-we-have-left-in-s%27pore

Singapore POFMA (2019). 'Protection From Online Falsehoods and Manipulation Act 2019'. Singapore Statutes online. https://sso.agc.gov.sg/Act/POFMA2019

Sippitt, A. (2020). 'Playing the long game: the impact of fact checkers on public figures, institutions, and the media'. *Fullfact*, 13 March 2020. https://fullfact.org/blog/2020/mar/long-game-impact-fact-checkers/

Smith, B. (2020a). 'When Facebook Is More Trustworthy Than the President'. *New York Times*, 15 March 2020. https://www.nytimes.com/2020/03/15/business/media/coronavirus-facebook-twitter-social-media.html

Smith, A. (2020b). 'Zuckerberg Criticises Twitter After It Fact-Checks Trump Tweets, Saying It Shouldn't Be 'Arbiter Of Truth''. *Independent*, 28 May 2020. https://www.independent.co.uk/life-style/gadgets-and-tech/news/trump-twitter-fact-check-mark-zuckerberg-jack-dorsey-a9536051.html?amp

Smith, A. & Shabad, R. (2020). 'Trump signs executive order aimed at social media companies after fuming over fact-check'. *NBC News*, 28 May 2020. https://www.nbcnews.com/politics/white-house/angry-over-how-social-media-platforms-are-treating-him-trump-n1216401

Smyrnaios, N., Chauvet, S. & Marty, E. (2017). 'The Impact of CrossCheck on Journalists & the Audience.' *First Draft*. https://firstdraftnews.org/wp-content/uploads/2017/11/Crosscheck_rapport_EN_1129.pdf/

Snapchat. (2017). 'Introducing the new Snapchat'. *Snap Inc*, 29 November 2017. https://www.snap.com/en-US/news/post/introducing-the-new-snapchat/

Snapchat. (2020). 'Coronavirus: How Snapchatters are reacting to the news and staying informed'. 17 March 2020. https://forbusiness.snapchat.com/blog/coronavirus-how-snapchatters-are-reacting-to-the-news-and-staying-informed

Soares, I. (2017) 'The Macedonia Story'. *CNN*. https://money.cnn.com/interactive/media/the-macedonia-story/

Solaiman, I, Clark, J. & Brundage, M. (2019). 'GPT-2: 1.5B Release'. *Open AI*, 5 November 2019. https://openai.com/blog/gpt-2-1-5b-release/

Soltani, A. (2018). 'Oral evidence: Disinformation and 'fake news'. HC 363. Questions 4274 – 4382. Published on 27 November 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture- media-and-sport-committee/disinformation-and-fake-news/oral/92924.html

Soundararajan, T., Kumar, A., Nair, P., Greely, J. (2019). 'Facebook India Towards the Tipping Point of Violence: Caste And Religious Hate Speech'. *Equality Labs*. https://www.equalitylabs.org/facebookindiareport

South African Government. (2020). Amendments to the 'Disaster Management Act, 2002, 18 March 2020. https://www.gov.za/sites/default/files/gcis_document/202003/43107gon318.pdf

Sovuthy, K. (2018). 'Government to tackle fake news frenzy'. *Khmer Times*, 05 July 2018. https://www.khmertimeskh.com/508265/government-to-tackle-fake-news-frenzy/

Spencer, S. (2019). 'An update on our political ads policy'. *Google Ads*, 20 November 2019. https://www.blog.google/technology/ads/update-our-political-ads-policy/

Spiegel, E. (2017). 'How Snapchat is separating social from media'. *Axios*, 29 November 2017. https://www.axios.com/how-snapchat-is-separating-social-from-media-2513315946.html

Spinney, L. (2019). 'Fighting Ebola is hard. In Congo, fake news makes it harder'. *AAAS Science Mag*, 14 January 2019. https://www.sciencemag.org/news/2019/01/fighting-ebola-hard-congo-fake-news-makes-it-harder

Spivak, C. (2010). 'The Fact-Checking Explosion. American Journalism Review'. University of Maryland. Winter 2010

Spring, M. (2020). 'Coronavirus: The seven types of people who start and spread viral misinformation'. *BBC News*, 04 May 2020. https://www.bbc.co.uk/news/blogs-trending-52474347?utm_source=First+Draft+Subscribers&utm_campaign=41894fb736-EMAIL_CAMPAIGN_2020_03_26_08_17_COPY_01&utm_medium=email&utm_term=0_2f24949eb0-41894fb736-265419209&mc_cid=41894fb736&mc_eid=a7e431ef92

Sri Lanka Brief. (2019). 'Sri Lanka govt to bring new laws to curb fake news & hate speech that jeopardize national harmony & national security'. 06 June 2019. https://srilankabrief.org/2019/06/sri-lanka-govt-to-bring-new-laws-to-curb-fake-news-hate-speech-that-jeopardize-national-harmony-national-security/

Srnicek, N. (2017). 'Platform Capitalism'. John Wiley & Sons, U.S

Stanley-Becker, I. (2020) 'Google Greenlights ads with "blatant disinformation" about voting by mail'. *The Washington Post*, 28 August 2020. https://www.washingtonpost.com/technology/2020/08/28/google-ads-mail-voting/?utm_source=twitter&utm_medium=social&utm_campaign=wp_main

Stanton, D. (2019). 'Advancing research on fake audio detection'. *Google News Initiative*, 31 January 2019. https://blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/

Staff, T. (2019). 'Election judge bars anonymous internet ads despite Likud objection'. *The Times of Israel*, 28 February, 2019. https://www.timesofisrael.com/election-judge-bars-anonymous-internet-adds-despite-likud-objection/

Starbird, K. (2017). 'Information Wars: A Window into the Alternative Media Ecosystem'. *In Proceedings of the 11th International AAAI Conference on Web and Social Media* (*ICWSM-17*), 15 March 2017. https://medium.com/hci-design-at-uw/information-wars-a-window-into-the-alternative-media-ecosystem-a1347f32fd8f

Statt, N. (2020). 'YouTube is a $15 billion-a-year business, Google reveals for the first time'. *The Verge*, 03 February 2020. https://www.theverge.com/2020/2/3/21121207/youtube-google-alphabet-earnings-revenue-first-time-reveal-q4-2019

Stecklow, S. (2018). 'Why Facebook is losing the war on hate speech in Myanmar'. *Reuters*, 15 August 2018. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Stewart, E. (2019). 'Facebook is refusing to take down a Trump ad making false claims about Joe Biden'. *Vox*, 9 October 2019. https://www.vox.com/policy-and-politics/2019/10/9/20906612/trump-campaign-ad-joe-biden-ukraine-facebook

Stolton, S. (2020) 'EU Rapid Alert System used amid coronavirus disinformation campaign'. EurActiv, 04 March 2020. https://www.euractiv.com/section/digital/news/eu-alert-triggered-after-coronavirus-disinformation-campaign/

Storm, H. (2020). 'Media ethics, safety and mental health: reporting in the time of Covid-19'. *EJN*, 12 March 2020. https://ethicaljournalismnetwork.org/media-ethics-safety-and-mental-health-reporting-in-the-time-of-covid-19

Strobelt, H. & Gehrmann, S. (2019). 'Catching a Unicorn with GLTR: A tool to detect automatically generated text'. http://gltr.io/

Suárez, E. (2020). '"I think removing content by Donald Trump will always have to meet a very high bar". *Reuters Institute for the study of Journalism, University of Oxford*, 19 May 2020. https://reutersinstitute.politics.ox.ac.uk/risj-review/i-think-removing-content-donald-trump-will-always-have-meet-very-high-bar

Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016). 'The DARPA Twitter Bot Challenge'. *Computer* Volume 49 Issue 6: pp.38–46. https://doi.org/10.1109/MC.2016.183

Swart, J., Peters, C., & Broersma, M. (2017). 'Navigating cross-media news use: Media repertoires and the value of news in everyday life'. *Journalism Studies*, 18(11), 1343-1362. https://doi.org/10.1080/1461670X.2015.1129285

Swedish Government (2018). En ny myndighet för psykologiskt försvar. Dir 2018-80. http://www.sou.gov.se/wp-content/uploads/2019/05/Direktiv-2018-80-En-ny-myndighet-för-psykologiskt-försvar.pdf

Swire, B. & Ecker, U.K.H. (2018). 'Misinformation and its Correction: Cognitive Mechanisms and Recommendations for Mass Communication'. In Southwell, B, Thorson, E.A. & Sheble, L (Eds.), '*Misinformation and Mass Audiences*'. Austin: University of Texas Press.

Sydell, L. (2016). 'We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned'. *NPR*, 23 November 2016. http://www.cpr.org/news/npr-story/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs

Symonds, T. (2018). 'The Skripals and the Salisbury poisoning: What happened next?'. *BBC News*, 28 December 2018. https://www.bbc.com/news/uk-46538255

Szunyogh, B. (1955). 'Psychological warfare; an introduction to ideological propaganda and the techniques of psychological warfare'. *William-Frederick Press*, United States. p. 13. Retrieved 2015-02-11.

Taibbi, M. (2019). 'YouTube, Facebook Purges Are More Extensive Than You Think'. Rolling Stone, 07 June 2019. https://www.rollingstone.com/politics/politics-features/youtube-facebook-purges-journalists-845790/

Tameez, H. (2020). 'YouTube's algorithm is pushing climate misinformation videos, and their creators are profiting from it'. *Nieman Lab*, 16 January 2020. https://www.niemanlab.org/2020/01/youtubes-algorithm-is-pushing-climate-misinformation-videos-and-their-creators-are-profiting-from-it/

Tanakasempipat, P. (2019a). 'Thailand passes internet security law decried as 'cyber martial law'. *Reuters*. 28 February 2019. https://www.reuters.com/article/us-thailand-cyber/thailand-passes-internet-security-law-decried-as-cyber-martial-law-idUSKCN1QH1OB

Tanakasempipat, P. (2019b). 'Thailand unveils 'anti-fake news' center to police the internet'. *Reuters*, 1 November 2019. https://www.reuters.com/article/us-thailand-fakenews/thailand-unveils-anti-fake-news-center-to-police-the-internet-idUSKBN1XB48O

Tapsell, R. (2019). 'Arrests for political hoax news in Indonesia'. *East Asia Forum*. 08 November 2019. https://www.eastasiaforum.org/2019/11/08/arrests-for-political-hoax-news-in-indonesia/

Tarbush, B., Teytelboym, A. (2012). 'Homophily in Online Social Networks'. *In*: *Goldberg P.W.* (*eds*) *Internet and Network Economics*. WINE 2012. Lecture Notes in Computer Science, vol 7695. Springer. https://doi.org/10.1007/978-3-642-35311-6_40

Tardáguila, C. (2019). 'Fact-checkers launch Instagram, WhatsApp and Telegram stickers to gently warn about false news'. *Poynter*, 05 June 2019. https://www.poynter.org/fact-checking/2019/fact-checkers-launch-instagram-whatsapp-and-telegram-stickers-to-gently-warn-about-false-news/

Tardáguila, C. (2020). 'Coronavirus: Fact-checkers from 30 countries are fighting 3 waves of misinformation'. *Poynter*, 28 January 2020. https://www.poynter.org/fact-checking/2020/coronavirus-fact-checkers-from-30-countries-are-fighting-3-waves-of-misinformation/

Tardáguila, C., Benevenuto, F. & Ortellado, P. (2018). 'Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It.' *The New York Times*, 17 October 2018. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html

Taylor, L. (2020). 'The coronavirus story is unfathomably large. We must get the reporting right'. *The Guardian*, 21 March 2020. https://www.theguardian.com/media/2020/mar/22/the-coronavirus-story-is-unfathomably-large-we-must-get-the-reporting-right

@Team Trump. (2019). '@facebook wants to take important tools away from us for 2020.' *Twitter*, 20 November 2019. https://twitter.com/TeamTrump/status/1197217636662337536?s=20

Tennessee General Assembly Fiscal Review Committee. (2020). 'Fiscal Note HJR 779'. 04 February 2020. http://www.capitol.tn.gov/Bills/111/Fiscal/HJR0779.pdf

Tennessee Legislature. (2020). 'House Joint Resolution 779'. 03 October 2019. http://www.capitol.tn.gov/Bills/111/Bill/HJR0779.pdf

Teyssou, D., Leung, J.M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., & Mezaris, V. (2017). 'The InVID Plug-in: Web Video Verification on the Browser. *In Proceedings of the First International Workshop on Multimedia Verification*'. (*pp. 23-30*). *ACM*, October 2017. https://www.researchgate.net/publication/320570485_The_InVID_Plug-in_Web_Video_Verification_on_the_Browser

Thailand Computer Crime Act. (2017). https://en.wikisource.org/wiki/Translation:Computer_Crimes_Act_(No._2)_2017

Thamm, M. (2019). 'Navigating your way in a world filled with untruths'. *Daily Maverick*, 26 September 2019. https://www.dailymaverick.co.za/article/2019-09-26-navigating-your-way-in-a-world-filled-with-untruths/

Theisen, W., Brogan, J., Bilo Thomas, P., Moreira, D., Phoa, P., Weninger, T. & Scheirer, W. (2020). 'Automatic Discovery of Political Meme Genres with Diverse Appearances' *ArXiv e-print*, January 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200106122T/abstract

The Jakarta Post. (2016). 'ITE Law draft revision passed into law'. 27 October 2016. https://www.thejakartapost.com/news/2016/10/27/ite-law-draft-revision-passed-into-law.html

The Jakarta Post. (2019). 'Stop Hoax Indonesia program to educate internet users in 17 cities'. 11 August 2019. https://www.thejakartapost.com/life/2019/08/10/stop-hoax-indonesia-program-to-educate-internet-users-in-17-cities.html

Theocharis, Y., Barberà, P., Fazekas, Z., Popa, S. A.,& Parnet, O. (2016). 'A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when interacting with party candidates'. Journal of Communication 66(6): 1007–1031, 28 October 2016. https://onlinelibrary.wiley.com/doi/full/10.1111/jcom.12259

Thomas, P. (2020). 'Coronavirus: Media pundits call for networks to stop airing Trump's 'misleading' briefings', *Independent*, 22 March 2020. https://www.independent.co.uk/news/world/americas/us-politics/coronavirus-trump-white-house-briefing-rachel-maddow-margaret-sullivan-jennifer-senior-a9416741.html

Thompson, S. (2020). 'Fact Check: As Of March 7, 2020, Tanzania And Zambia Had NOT Confirmed First Cases Of Coronavirus'. *Lead Stories*, 07 March 2020. https://leadstories.com/hoax-alert/2020/03/fact-check-tanzania-confirms-first-case-of-coronavirus---mcm.html

Thompson, T. (2019). 'Countering Russian disinformation the Baltic nations' way'. *The Conversation*, 9 January, 2019. https://theconversation.com/countering-russian-disinformation-the-baltic-nations-way-109366

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). 'FEVER: a large-scale dataset for Fact Extraction and VERification'. Cornell University, 14 March 2018. https://arxiv.org/abs/1803.05355

Tidy, J. & Schraer, R. (2019). 'General election 2019: Ads are 'indecent, dishonest and untruthful'. *BBC*, 17 December 2019. https://www.bbc.co.uk/news/technology-50726500

TIMEP. (2019). 'TIMEP Brief: The Law Regulating the Press, Media, and the Supreme Council for Media Regulation'. 15 May, 2019. https://timep.org/reports-briefings/timep-brief-the-law-regulating-the-press-media-and-the-supreme-council-for-media-regulation/

TNN. (2020). 'How India Became The Global Leader in Internet Shutdowns'. *Times of India*, 08 January 2020. https://timesofindia.indiatimes.com/india/how-india-became-the-global-leader-in-internet-shutdowns/articleshow/72886376.cms?from=mdr

Tracy, M. (2020). 'News Media Outlets Have Been Ravaged by the Pandemic'. *The New York Times*, 10 May 2020. https://www.nytimes.com/2020/04/10/business/media/news-media-coronavirus-jobs.html

Trend Micro. (2020). 'Developing Story: COVID-19 Used in Malicious Campaigns'. 24 April 2020. https://www.trendmicro.com/vinfo/fr/security/news/cybercrime-and-digital-threats/coronavirus-used-in-spam-malware-file-names-and-malicious-domains

Troop, W. (2017). 'This Italian politician wants kids to become 'fake news hunters'. *The World*, 31 October 2017. https://www.pri.org/stories/2017-10-31/italian-politician-wants-kids-become-fake-news-hunters

Tsai, C.-H. & Brusilovsky, P. (2018). 'Beyond the Ranked List: User-Driven Exploration and Diversification of Social Recommendation'. *In Proceedings of the 23rd International Conference on Intelligent User Interfaces. ACM*. https://pennstate.pure.elsevier.com/en/publications/beyond-the-ranked-list-user-driven-exploration-and-diversificatio

Turvill, W. (2020). 'Going viral: Twitter and Facebook failing to contain Covid-19 misinformation'. *PressGazette*, 31 March 2020. https://www.pressgazette.co.uk/the-fake-news-epidemic-twitter-and-facebook-failing-to-contain-covid-19-misinformation/

Twitter. (2018). Oral Evidence for Lords Communications Committee – 'The Internet: To Regulate or Not To Regulate?' https://parliamentlive.tv/Event/Index/2cd62e7a-d3cf-4605-8d39-4fbaa0adaa76#player-tabs

Twitter. (2019a). 'New disclosures to our archive of state-backed information operations'. 20 December 2019. https://blog.twitter.com/en_us/topics/company/2019/new-disclosures-to-our-archive-of-state-backed-information-operations.html

Twitter. (2019b). 'Twitter Progress Report: Code of Practice against Disinformation'. https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019

UK Commons Select Committee on Digital, Culture, Media and Sport. (2018). *Parliamentarians from across the World to Question Richard Allan of Facebook, and the Information Commissioner at Inaugural Hearing of 'International Grand Committee' on Disinformation and 'Fake News'*, Press Release, 23 November 2018. https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/grand-committee-evidence-17-19/

UK Commons Select Committee on Digital, Culture, Media and Sport. (2019). *Disinformation and 'Fake News': Final Report*. London: House of Commons. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

UK Department for Digital, Culture, Media and Sport (DCMS). (2018b). 'Oral evidence : Fake News – 08 February 2018' (George Washington University, Washington DC), HC 363. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culturemedia-and-sport-committee/fake-news/oral/78195.html

UK Department for Digital, Culture, Media and Sport (DCMS). (2018c). 'Disinformation and 'fake news': Interim Report: Government Response to the Committee's Fifth Report of Session 2017–19. HC 1630. Published on 23 October 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1630/1630.pdf

UK Department for Digital, Culture, Media and Sport and UK Home Office (DCMS). (2019). 'Online Harms White Paper'. 12 February 2019. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

UK Department for Digital, Culture, Media and Sport and UK Home Office (DCMS). (2020). 'Online Harms White Paper – Initial consultation response'. 12 February 2020. https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

UK DCMS (Department for Digital, Culture, Media and Sport), UK Home Office, The Rt Hon Matt Hancock MP & The Rt Hon Sajid Javid MP. (2018a). 'New laws to make social media safer'. Press Release, 20 May 2018. https://www.gov.uk/government/news/new-laws-to-make-social-media-safer

UK DCMS. (Department for Digital, Culture, Media and Sport) & The Rt Hon Matt Hancock MP. (2018b). Policy Paper – 'Digital Charter', 25 January 2018. https://www.gov.uk/government/publications/digital-charter

UK Delegation to the OSCE. (2020). 'Joint statement on safety of journalists and access to information during the COVID-19 crisis'. 15 April 2020. https://www.gov.uk/government/news/joint-statement-on-safety-of-journalists-and-access-to-information-during-the-covid-19-crisis--2

UK Department for International Development. (2020). 'UK aid to tackle global spread of coronavirus 'fake news''. 12 March 2020. https://www.gov.uk/government/news/uk-aid-to-tackle-global-spread-of-coronavirus-fake-news

UK Government. (2019). 'Conflict, Stability and Security Fund: programme summaries for Eastern Europe, Central Asia and Western Balkans 2019 to 2020'. 05 November 2019. https://www.gov.uk/government/publications/conflict-stability-and-security-fund-programme-summaries-for-eastern-europe-central-asia-and-western-balkans-2019-to-2020

UK Government. (2020). 'UK aid to tackle global spread of coronavirus 'fake news'. Press Release, 12 March 2020. https://www.gov.uk/government/news/uk-aid-to-tackle-global-spread-of-coronavirus-fake-news

UK House of Commons Foreign Affairs Committee. (2019). '"Media freedom is under attack": The FCO's defence of an endangered liberty'. 04 September 2019. https://publications.parliament.uk/pa/cm201719/cmselect/cmfaff/1920/1920.pdf

UK ICO. (2018). *ICO Issues Maximum £500,000 Fine to Facebook for Failing to Protect Users' Personal Information*. Press Release, 25 October 2018. https://ico.org.uk/facebook-fine-20181025

UK MOD (Ministry of Defence), Defence Science and Technology Laboratory, Guto Bebb, and The Rt Hon Gavin Williamson CBE MP. (2018). 'Flagship AI Lab announced as Defence Secretary hosts first meet between British and American defence innovators'. Press Release, 22 May 2018. https://www.gov.uk/government/news/flagship-ai-lab-announced-as-defence-secretary-hosts-first-meet-between-british-and-american-defence-innovators

UK Parliament, Subcommittee on Online Harms and Disinformation. (2020). 'Disinformation and misinformation on social media about COVID-19 to be investigated'. 26 March 2020. https://committees.parliament.uk/committee/438/subcommittee-on-online-harms-and-disinformation/news/145761/disinformation-and-misinformation-on-social-media-about-covid19-to-be-investigated/

UN. (1948). 'Universal Declaration of Human Rights'. General Assembly Resolution 217A, 10 December 1948. https://www.un.org/en/universal-declaration-human-rights/

UN. (2019). United Nations Strategy and Plan of Action on Hate Speech. May 2019. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

UN Africa Renewal. (2020). 'Mauritius, Senegal, South Africa among authors of global call against COVID-19 'infodemic'. 22 June 2020. https://www.un.org/africarenewal/news/coronavirus/cross-regional-statement-%E2%80%9Cinfodemic%E2%80%9D-context-covid-19

UN Department of Global Communications. (2020). '5 ways the UN is fighting 'infodemic' of misinformation'. 30 April 2020. https://www.un.org/en/un-coronavirus-communications-team/five-ways-united-nations-fighting-%E2%80%98infodemic%E2%80%99-misinformation

UN Human Rights. (2018). 'UN experts call on India to protect journalist Rana Ayyub from online hate campaign'. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23126&LangID=E

UN Human Rights. (2020a). 'COVID-19: Governments must promote and protect access to and free flow of information during pandemic – International experts. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25729&LangID=E

UN Human Rights. (2020b). 'Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development'. 14 July 2020. A/HRC/44/L.18/Rev.1 https://undocs.org/A/HRC/44/L.18/Rev.1

UN News. (2020). 'Hatred going viral in 'dangerous epidemic of misinformation' during COVID-19 pandemic'. 14 April 2020. https://news.un.org/en/story/2020/04/1061682

UN Secretary General. (2020). 'Secretary-General's remarks to High-Level Dialogue on Press Freedom and Tackling Disinformation in the COVID-19 Context [bilingual, as delivered; scroll down for English and French versions]'. 04 May 2020. https://www.un.org/sg/en/content/sg/statement/2020-05-04/secretary-generals-remarks-high-level-dialogue-press-freedom-and-tackling-disinformation-the-covid-19-context-bilingual-delivered-scroll-down-for-english-and-french

UN Special Rapporteur on Freedom of Opinion and Expression et. al. (2017). 'Joint Declaration on Freedom of Expression and "Fake News," Disinformation and Propaganda'. UN Document FOM.GAL/3/17, https://www.osce.org/fom/302796?download=true

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018a). Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression. A/73/348. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018b). Report to the United Nations Human Rights Council on A Human Rights Approach to Platform Content Regulation, A/HRC/38/35, https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

UNESCO. (2016). 'UNESCO launches Countering Online Hate Speech publication'. https://unesdoc.unesco.org/ark:/48223/pf0000233231

UNESCO. (2019). 'Elections and media in digital times'. https://en.unesco.org/sites/default/files/in_focus_world_trends_report_2019_brochure_final_2.pdf

UNESCO. (2020). 'Journalism, press freedom and COVID-19'. https://en.unesco.org/sites/default/files/unesco_covid_brief_en.pdf

UNI Global Union. (2020). 'Solidarity: Fighting the COVID-19 Crisis'. https://www.uniglobalunion.org/sites/default/files/files/news/20_03_25_en_uni_mei_report_on_response_to_covid-19_in_the_media_enteratinment_sector.pdf

United Nations. (1994). 'Professional Training Series No. 2. Human Rights and Elections. A Handbook on the Legal, Technical and Human Rights Aspects of Elections. https://eos.cartercenter.org/uploads/document_file/path/8/training2enTCCoptimized.pdf

US Senate Select Committee on Intelligence. (2018). 'New Reports Shed Light on Internet Research Agency's Social Media Tactics'. Press Release, 17 December 2017. https://www.intelligence.senate.gov/press/new-reports-shed-light-internet-research-agency%E2%80%99s-social-media-tactics

US Senate Select Committee on Intelligence. (2019a). 'Russian Active Measures, Campaigns and Interference In the 2016 U.S. Election'. Volume 1, July 2019. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf

US Senate Select Committee on Intelligence. (2019b). 'Russian Active Measures, Campaigns and Interference In the 2016 U.S. Election'. Volume 2, October 2019. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

US Senate Select Committee on Intelligence. (2019c). 'Russian Active Measures, Campaigns and Interference In the 2016 U.S. Election'. Volume 3, February 2020. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume3.pdf

US Senate Select Committee on Intelligence. (2019d). 'Senate Intel Committee Releases Bipartisan Report on Russia's Use of Social Media'. Press Release, 08 October 2019. https://www.intelligence.senate.gov/press/senate-intel-committee-releases-bipartisan-report-russia's-use-social-media

Vaccari, C. & Chadwick, A. (2020). 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News'. *Social Media + Society*, 19 February 2020. https://doi.org/10.1177/2056305120903408

Van Den Berg, E. & Snelderwaard, T. (2019). 'Nu.nl stopt met bestrijding nepnieuws voor Facebook na terugdraaien factchecks PVV en FvD'. *NPO3*, 26 November 2019. https://www.npo3.nl/brandpuntplus/nu-nl-stopt-met-bestrijding-nepnieuws-voor-facebook-we-willen-ook-de-pvv-kunnen-factchecken

Van Dieman, E. (2019). '70 complaints of 'digital disinformation' over past month - Electoral Commission'. *News 24*, 06 May 2019. https://www.news24.com/elections/news/70-complaints-of-digital-disinformation-over-past-month-electoral-commission-20190506

Van Dijck, J., Poell, T. & de Waal, M. (2018). 'The Platform Society: Public Values in a Connective World'. Oxford University Press, 2018.

Varol, O., Ferrara, E., Davis, C.A., Menczer, F. & Flammini, A. (2017). 'Online Human-Bot Interactions: Detection, Estimation, and Characterization'. In Intl. AAAI Conf. on Web and Social Media (ICWSM), 09 March 2017. https://arxiv.org/abs/1703.03107

Verberne, S. (2018). 'Explainable IR for Personalizing Professional Search'. In Joint Proceedings of the 1st International Workshop on Professional Search (ProfS2018), co-located with SIGIR. http://webcache.googleusercontent.com/search?q=cache:GeXr_yAHVGYJ:ceur-ws.org/Vol-2127/paper4-profs.pdf+&cd=1&hl=en&ct=clnk&gl=uk&client=firefox-b-e

Verdoliva, L. (2020). 'Media Forensics and DeepFakes: an overview'. https://arxiv.org/abs/2001.06564

Vicol, D-O. (2020). 'What makes us believe a false claim? Age, education, and cognitive biases all play a part'. *Fullfact*, 28 February 2020. https://fullfact.org/blog/2020/feb/what-makes-us-believe-false-claim/

Viner, K. (2017). 'A mission for journalism in a time of crisis'. *The Guardian*, 17 November 2017. https://www.theguardian.com/news/2017/nov/16/a-mission-for-journalism-in-a-time-of-crisis

Vishwanath, A. (2020). 'Explained: The laws being used to suspend Internet, and what SC laid down'. *The Indian Express*, 11 January 2020. https://indianexpress.com/article/explained/kashmir-supreme-court-internet-shutdown-sec-144-how-to-read-judgment-6209676/

Vlachos, A., & Riedel, S. (2015). 'Identification and Verification of Simple Claims about Statistical Properties'. In volume 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', 17-21 September 2015, Lisbon, Portugal. Association for Computational Linguistics, 2596-2601. https://www.aclweb.org/anthology/D15-1312/

Vosoughi, S., Roy, D. & Aral, S. (2018). 'The spread of true and false news online'. *Science*: Vol. 359, Issue 6380, pp. 1146-1151

Waddell, K. (2019) 'A digital breadcrumb trail for deepfakes.' *AXIOS*, 12 July 2019. https://www.axios.com/deepfake-authentication-privacy-5fa05902-41eb-40a7-8850-5450bcad0475.html

Walker, S., Mercea, D. & Bastos, M. (2019). 'The disinformation landscape and the lockdown of social platforms'. https://doi.org/10.1080/1369118X.2019.1648536

Wang, A.H. (2010). 'Don't follow me: Spam detection in Twitter'. *In Proceedings of the International Conference on Security and Cryptography* (*SECRYPT 2010*). https://ieeexplore.ieee.org/abstract/document/5741690

WAN-IFRA. (2016). 'The 2016 Global Report on Online Commenting: Executive Summary'. 17 October 2016. https://blog.wan-ifra.org/2016/10/17/the-2016-global-report-on-online-commenting-executive-summary

WARC. (2018). '65% of digital media will be programmatic in 2019'. 19 November, 2018. https://www.warc.com/newsandopinion/news/65_of_digital_media_will_be_programmatic_in_2019/41341

Wardle, C. (2017a). 'Fake News. it's Complicated'. *First Draft*, 16 February 2017. https://firstdraftnews.org/fake-news-complicated/,

Wardle, C. (2017b). 'Foreword in The Impact of CrossCheck on Journalists & the Audience'. *First Draft*. https://firstdraftnews.org/wp-content/uploads/2017/11/Crosscheck_rapport_EN_1129.pdf/

Wardle, C. & Derakhshan, H. (2017). 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making'. Council of Europe report DGI, 09 October 2017. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Warner, M. R. (2017). 'The Honest Ads Act'. US Senator of Virginia, US, May 2019. https://www.warner.senate.gov/public/index.cfm/the-honest-ads-act

Wasserman, H. (2020). 'Laughter in the time of a pandemic: why South Africans are joking about coronavirus'. *The Conversation*, 15 March 2020. https://theconversation.com/laughter-in-the-time-of-a-pandemic-why-south-africans-are-joking-about-coronavirus-133528

Wasserman, H. & Madrid-Morales, D. (2018). 'Study sheds light on scourge of "fake" news in Africa'. The Conversation, 21 November 2018. https://theconversation.com/study-sheds-light-on-scourge-of-fake-news-in-africa-106946

WebWire. (2020). 'NEW REPORT: YouTube found promoting climate denial to millions'. 20 January 2020. https://www.webwire.com/ViewPressRel.asp?aId=253667

Weedon, J., Nuland, W. & Stamos, A. (2017). 'Information Operations and Facebook'. *Facebook*, 27 April 2017, https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

Weinberger, D. (2009). 'Transparency is the new objectivity'. https://www.hyperorg.com/blogger/2009/07/19/transparency-is-the-new-objectivity/

WhatsApp. (2019a). 'More changes to forwarding'. *WhatsApp Blog*, 21 January 2019. https://blog.whatsapp.com/more-changes-to-forwarding

WhatsApp. (2019b). 'Stopping Abuse: How WhatsApp Fights Bulk Messaging and Automated Behavior'. https://scontent.whatsapp.net/v/t61/69510151_652112781951150_6923638360331596993_n.pdf/Stopping-Abuse-white-paper.pdf?_nc_sid=2fbf2a&_nc_ohc=hwbXZHmgl3oAX9FN2oR&_nc_ht=scontent.whatsapp.net&oh=c5105a645efd83ac2e78d5e3e33247c3&oe=5E61E6EE

WHO. (2018). 'Risk Communication and Community Engagement (RCCE) Considerations: Ebola Response in the Democratic Republic of the Congo'. May 2018. https://www.afro.who.int/publications/risk-communication-and-community-engagement-rcce-considerations-ebola-response

WHO. (2020). Coronavirus, 13 February 2020. Press Conference Transcript: https://www.who.int/docs/default-source/coronaviruse/transcripts/who-audio-emergencies-coronavirus-full-press-conference-13feb2020-final.pdf?sfvrsn=b5435aa2_2

Who Targets Me. (2019). 'Despite protestations from the @bbc, the @Conservatives haven't stopped running the ad.' *Twitter*, 30 November 2019. https://twitter.com/WhoTargetsMe/status/1200729905408987136

Wilding, D., Fray, P., Molitorisz, S. & McKewon, E. (2018). 'The Impact of Digital Platforms on News and Journalistic Content'. University of Technology Sydney. https://www.accc.gov.au/system/files/ACCC%20commissioned%20report%20-%20The%20impact%20of%20digital%20platforms%20on%20news%20and%20journalistic%20content%2C%20Centre%20for%20Media%20Transition%20%282%29.pdf

Wilner, T. (2018). 'We can probably measure media bias. But do we want to?' *Columbia Journalism Review*, 9 January 2018. https://www.cjr.org/innovations/measure-media-bias-partisan.php

Witness Media Lab. (2019). 'Ticks or it didn't happen: Confronting key dilemmas in authenticity infrastructure for multimedia'. December 2019. https://lab.witness.org/ticks-or-it-didnt-happen/

Wong, J. C. (2019a). 'Google latest tech giant to crack down on political ads as pressure on Facebook grows'. *The Guardian*, 21 November 2019. https://www.theguardian.com/technology/2019/nov/20/google-political-ad-policy-facebook-twitter

Wong, J. C. (2019b). 'Sri Lankans fear violence over Facebook fake news ahead of election'. *The Guardian*, 12 November 2019. https://www.theguardian.com/world/2019/nov/11/facebook-sri-lanka-election-fake-news

Wong, J. C. (2020a). 'Will Facebook's new oversight board be a radical shift or a reputational shield?'. *The Guardian*, 07 May 2020. https://www.theguardian.com/technology/2020/may/07/will-facebooks-new-oversight-board-be-a-radical-shift-or-a-reputational-shield

Wong, J. C. (2020b). 'Facebook removes Trump post over false Covid-19 claim for first time'. *The Guardian*, 06 August 2020. https://www.theguardian.com/technology/2020/aug/05/facebook-donald-trump-post-removed-covid-19

Wood, A. K. & Ravel, A.M. (2018). 'Fool me once: regulating "fake news" and other online advertising'. *Southern California Law Review*. Vol. 91. Pages 1223-1278. https://southerncalifornialawreview.com/wp-content/uploads/2018/10/91_6_1223.pdf

Woolley, S.C. & Howard, P.N. (2016). 'Automation, Algorithms, and Politics: Political Communication, Computational Propaganda, and Autonomous Agents. *International Journal of Communication*, 10 October 2016. http://ijoc.org/index.php/ijoc/article/view/6298/1809

World Summit Working Group. (2005). 'Report of the Working Group on Internet Governance'. United Nations & World Summit, June 2005. https://www.wgig.org/docs/WGIGREPORT.pdf

Wu, T. (2017). 'Blind Spot: The Attention Economy and the Law'. *Antitrust Law Journal, Forthcoming*, 26 March 2017. https://ssrn.com/abstract=2941094 or http://dx.doi.org/10.2139/ssrn.2941094

Wulczyn, E., Thain, N., & Dixon, L. (2017). 'Ex machina: Personal attacks seen at scale'. *In Proceedings of the 26th International Conference on World Wide Web*, *1391–1399*, 25 February 2017. https://arxiv.org/abs/1610.08914

Xinhua. (2019). 'Consensus issued after fourth meeting of World Media Summit presidium in Shanghai'. 20 November 2019. http://www.xinhuanet.com/english/2019-11/20/c_138567780.htm

Yang, Z. (2018). '网信办公布新规 微博客服务提供者应建立健全辟谣机制'. *People's Daily*, 03 February 2018. http://www.xinhuanet.com/2018-02/03/c_1122362053.htm

Zacharia, J. (2019). 'Time is running out to to fight disinformation in 2020 election'. *San Francisco Chronicle*, 31 August 2019. https://www.sfchronicle.com/opinion/article/Time-is-running-out-to-to-fight-disinformation-in-14404399.php?psid=ms95Y

Zadrozny, B. (2019). 'Anti-vaccination groups still crowdfunding on Facebook despite crackdown'. *NBC News*, 11 October 2019. https://www.nbcnews.com/tech/internet/anti-vaccination-groups-still-crowdfunding-facebook-despite-crackdown-n1064981

Zampoglou, Z. Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., & Spangenberg, J. (2016). 'Web and Social Media Image Forensics for News Professionals'. *In Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 16 April 2016. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13206

Zarocostas, J. (2020). 'How to fight an infodemic'. The Lancet, 29 February 2020. https://doi.org/10.1016/S0140-6736(20)30461-X

Zeit. (2019). 'Deutsche Behörde verhängt Millionenstrafe gegen Facebook'. 02 July 2019. https://www.zeit.de/digital/datenschutz/2019-07/facebook-hasskommentare-fake-news-millionenstrafe-bussgeld-netzdg

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). 'Defending Against Neural Fake News'. 29 May 2019. http://arxiv.org/abs/1905.12616

Zharov, A. (2019). 'Russia to Set Up 'Fake News Database''. *The Moscow Times*, 16 May 2019. https://www.themoscowtimes.com/2019/05/16/russia-to-set-up-fake-news-database-a65613

Zilavy, T. (2018). 'What is Hashing and How Does It Work?'. Medium, 06 October 2018. https://medium.com/datadriveninvestor/what-is-hashing-and-how-does-it-work-7800f461a0de

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P. (2016). 'Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads'. PLOS ONE 11(3): e0150989. https://doi.org/10.1371/journal.pone.0150989

Zuboff, S. (2019). 'The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power'. New York: Public Affairs

Zuckerberg, M. (2016a). 'Announcement from Zuckerberg regarding the launch of a "third party verification" program'. *Facebook*, 18 November 2016. https://www.facebook.com/zuck/posts/10103269806149061

Zuckerberg, M. (2016b). 'I want to share some thoughts about Facebook and the election'. https://www.facebook.com/zuck/posts/i-want-to-share-some-thoughts-on-facebook-and-the-electionour-goal-is-to-give-ev/10103253901916271/

Zuckerberg, M. (2018). Testimony before Senate: "Why do you shift the burden to users to flag inappropriate content and make sure it's taken down?". *BuzzFeed*, 10 April 2018. https://twitter.com/BuzzFeedNews/status/983815663469948929

5Rights Foundation. (2019). 'Towards an Internet Safety Strategy'. *5Rights Foundation*, January 2019. https://5rightsfoundation.com/uploads/final-5rights-foundation-towards-an-internet-safety-strategy-january-2019.pdf

Appendix A

Inquiries, task forces and guidelines

With widespread misinformation becoming a growing concern, several countries have set up dedicated task forces and inquired to monitor and investigate disinformation campaigns. Such task forces have often been launched following disinformation campaigns perceived as a threat to the country's democratic integrity, or cyber-security. The primary scope of these governmental initiatives is the educational, with 24 out of 33 including a media literacy aspect. In addition, 17 initiatives in this category include fact-checking. It can be highlighted that out of the 33 countries which have set up such inquiries or task forces, 21 have an electoral-specific focus. Electoral-specific inquiries have the objective to investigate or prevent interference into legislative processes. Online disinformation being a relatively new phenomenon, most of the initiatives identified are recent and still susceptible to evolution, or might eventually lead to regulatory initiatives.

1. ASEAN Framework and Joint Declaration to Minimise the Harmful Effects of Fake News (2018)

The ASEAN Ministers responsible for Information (ASEAN AMRI, 2018) made a 'Joint Declaration on the Framework to Minimise the Harmful Effects of Fake News'. In it, they promote the sharing of best practices among ASEAN members and propose a framework building on four strands: education and awareness (digital literacy and own responsibility), detection and response (government monitoring and debunking of disinformation, media monitoring and fact-checking); community and ground-up participation (citizen and civil society detection of disinformation); norms and guidelines (depending on national context, laws, norms, guidelines to "empower and protect citizens" and "promote responsible generation and sharing of online information") (ASEAN AMRI, 2018).

2. Australia Electoral Integrity Assurance Taskforce (2019)

The Australian Electoral Commission (AEC, 2019a) formed a task force with other government agencies in the run-up to the federal elections in May 2019, in order to protect electoral integrity, with solutions focusing on authorisation of electoral communications, transparency in foreign influence, cyber-security. The Commission also ran a media literacy campaign 'Stop and Consider' (AEC, 2019b). Several government reports relating to media, in particular on the future of public interest journalism and the impact of digital platforms on media competition, have also highlighted the need for transparency and media literacy in combatting disinformation (Buckmaster & Wils, 2019).

3. Australia Parliament Joint Standing Committee on Electoral Matters: Democracy and Disinformation (2018)

Noting the presence of disinformation in the U.S. election in 2016, the Brexit referendum in the UK, and the evidence of disinformation on democratic processes internationally in 2018, the Australian Parliament Joint Standing Committee on Electoral Matters (2019) decided to incorporate "democracy and disinformation into its ongoing oversight of Australian electoral matters, with the following focus areas:

- the extent to which social media bots may have targeted Australian voters and political discourse in the past;
- the likely sources of social media manipulation within Australia and internationally;

- ways to address the spread of online 'fake news' during election campaigns; and
- measures to improve the media literacy of Australian voters."

4. Belgium Expert Group and Participatory Platform (2018)

The Belgian Minister for the Digital Agenda organised an expert group (Alaphilippe et.al., 2018a), a participatory platform (Monopinion.belgium.be)[488] and a citizen debate to discuss disinformation in 2018. As a result of these consultations, the federal government made available 1.5 million Euro to support media literacy initiatives that increase the transparency of digital ecosystems (e.g. advertising transparency) and the findability of high quality and diverse information (e.g. source indicators) (De Croo, 2018).

5. Brazil Superior Electoral Court (2018)

The Brazilian Superior Electoral Court (TSE) signed a memorandum of understanding with Google and Facebook to limit the spread of electoral disinformation in the run-up to the general elections in October 2018 (Alves, 2018).

6. Canada Parliamentary Committee Report on Democracy under Threat (2018)

In the wake of the Cambridge Analytica scandal, the Canadian House of Commons Standing Committee on Access to Information, Privacy and Ethics (2018) published a study on risks and solutions to democracy in the era of disinformation and data monopoly, with a broad range of recommendations from the application of privacy legislation to political parties and increasing the powers of the privacy commissioner, to social media platform regulation and cyber security, as well as support for digital literacy and research into the impact of disinformation.

7. Canada Government Digital Citizen Initiative (2019)

The Canadian government set up a 'Digital Citizen Initiative' prior to the federal elections in October 2019, "to build citizen resilience against online disinformation and building partnerships to support a healthy information ecosystem" (Canadian Government, 2019a; Canadian Government, 2019b). This initiative has resulted in substantial investment in responses (ranging from research to journalism education, and media literacy training) to the disinformation crisis.

8. Canada Government Critical Election Incident Public Protocol (2019)

The Critical Election Incident Public Protocol (CEIPP)[489] laid out a "simple, clear and impartial process" to guide the notification of Canadians about a threat to the integrity of the 2019 General Election. A group of experienced senior Canadian public servants comprised the CEIPP Panel (covering national security, foreign affairs, democratic governance and legal perspectives), with responsibility for determining whether the threshold for informing Canadians was constituted.

[488] https://monopinion.belgium.be/processes/stopfakenews
[489] https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html

9. Council of Europe Information Disorder Report (2017)

In October 2017, the Council of Europe published a Report 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making' (Wardle & Derakhshan, 2017). The report distinguishes between mis-information ("when false information is shared, but no harm is meant"), dis-information ("when false information is knowingly shared to cause harm") and 'mal-information' ("when genuine information is shared to cause harm, often by making public information designed to stay private"). It provides recommendations to tech companies, national governments, media, civil society, and education ministries. The Council of Europe also provides media and information literacy resources to tackle misinformation, disinformation and propaganda as part of its educational 'Free to Speak - Safe to Learn' programme.[490]

10. Denmark Elections Action Plan (2018)

In 2018, the Danish Ministry of Foreign Affairs (2018) announced an action plan to counter hostile foreign interference in Danish elections. The eleven initiatives range from an inter-governmental coordination task force and increased cyber-security preparedness, to the training of political parties and dialogue with media and social media companies on possible cooperation models to counter potential foreign influence attempts. The Danish government will also update the criminal code to protect Denmark against "the threat from influence campaigns launched by foreign intelligence services".

11. Estonia Cyber Defence League (2007-)

Estonia has historical experience of being targeted with propaganda and disinformation. After an intense cyber-attack in 2007, Estonia set up a Defence League Cyber Unit whose mission is to "protect Estonia's high-tech way of life, including protection of information infrastructure and supporting broader objectives of national defence".[491] The Unit includes volunteer patriotic citizens willing to contribute to the country's cyber security strategy. Since 2008, Estonia has also been home of the NATO's Cooperative Cyber Defence Centre of Excellence[492] (Thompson, 2019).

12. EU High-Level Expert Group, Code of Practice and Action Plan on Disinformation (2018)

In March 2018 a multi-stakeholder High Level Expert Group on Fake News and Online Disinformation (EU HLEG, 2018) issued recommendations on disinformation, focusing primarily on the role that social media platforms can play in supporting the media ecosystem, fact-checking and literacy efforts.

A follow-up European Commission Communication (European Commission, 2018a) included recommendations on media literacy and pluralism as well, although in a significantly reduced form. The Commission also added reflections on election processes and strategic communication. Importantly the Communication picked up on the HLEG's reflections on a transparent digital ecosystem and set up the EU multi stakeholder forum to develop an EU Code of Practice (European Commission, 2018b).

490 https://www.coe.int/en/web/campaign-free-to-speak-safe-to-learn/dealing-with-propaganda-misinformation-and-fake-news
491 https://www.kaitseliit.ee/en/cyber-unit
492 https://ccdcoe.org/

The resulting EU Code of Practice (European Commission, 2018c) focuses on (electoral) adverts, includes a short section on automated bots, and addresses the platforms' role in supporting/enabling literacy, fact-checking and research. The Code of Practice mainly recaps existing measures and does not aim to provide industry standards. One year onward, the implementation of the Code of Practice is under independent review (European Commission, 2019).

All this fits within the EU's wider Action Plan on Disinformation, which aims to build capacities and cooperation within the EU and among its member states (European Commission and High Representative, 2018). The European External Action Service also runs a website aiming to provide counter-narratives to disinformation.

13. India Social Media Platforms' Code of Ethics for the 2019 General Elections (2019)

Building on the report of the Committee on Section 126 of the Representation of the People Act, 1951 ('Sinha Committee Report') which made recommendations on maintaining campaign silence during last 48 hours prior to polling (ECI, 2019), the Social Media Platforms and Internet and Mobile Association of India (IAMAI, 2019) agreed upon a Voluntary Code of Ethics for the 2019 General Elections. A number of social media companies made commitments to run voter awareness campaigns, to "create a high priority dedicated reporting mechanism" for the Electoral Commission of India (ECI.). They also agreed to appoint dedicated teams to the elections, upon notification from the ECI to take down reported content within three hours during the 48-hour silence period prior to polling, to provide a mechanism for political advertisers to submit pre-certificates issued by the ECI and to act upon these requests expeditiously, and to facilitate transparency in paid political ad transparency (IAMAI, 2019).

14. Indonesia 'War Room' and 'Stop Hoax' Campaign (2019)

In the run-up to the presidential elections in April 2019, the Indonesian Ministry of Communication and Information Technology organised a 'war room' to detect and disable negative and violating content (Board, 2019).

The Ministry also teamed up with the Indonesian Anti-Slander Society, the Google News Initiative and several civil society organisations, to run a 'Stop Hoax' literacy campaign, which primarily targeted students and women working at home. The workshops, which ran from August to October 2019, aimed at helping participants to detect disinformation and hoaxes.

15. International Grand Committee on Disinformation and 'Fake News' (2018-)

The International Grand Committee on Disinformation and 'Fake News' found its origin in the UK Parliamentary inquiry into disinformation and fake news. The Grand Committee brings together parliamentarians and has so far heard evidence related to disinformation, big data, regulation of harmful content and electoral interference online. Three meetings have taken place in London (November 2018, hosted by the UK Parliamentary Digital, Culture, Media and Sport Committee), Ottawa (May 2019, hosted by Canadian Parliamentary Standing Committee on Access to Information, Privacy and Ethics) and Dublin (November 2019, hosted by the Irish Joint Oireachtas Committee on Communication, Climate Action and Environment).

16. Ireland Interdepartmental Group on Security of the Election Process and Disinformation (2017-)

In December 2017, the Irish government established an interdepartmental group to assess threats to the Irish electoral process (IDG). In the group's first report, published in July 2018, the main finding was that "risks to the electoral process in Ireland are relatively low but that the spread of disinformation online and the risk of cyber-attacks on the electoral system pose more substantial risks" (Irish IDG, 2019). The report outlined seven recommendations, including establishing an Electoral Commission, continuing media literacy initiatives and enhancing cyber security measures. In November 2019, the Irish government introduced a legislative proposal to regulate the transparency of online paid political advertising within election periods, in line with another of the interdepartmental group's recommendations (Irish Department of the Taoiseach, 2019).

17. Italy 'Enough-With-the-Hoaxes' Campaign and 'Red Button' Portal (2018-)

In October 2017 the Italian Ministry of Education (2017) announced an 'enough-with-the-hoaxes' (Bastabufale.it)[493] media literacy campaign for primary and secondary schools. Before the general elections in March 2018, a 'red button' portal[494] was also launched, where citizens could report disinformation to a special cyber police unit. The police unit would investigate the content, help citizens report disinformation to social media platforms, and in case of defamatory or otherwise illegal content, file a lawsuit (la Cour, 2019). The Italian communications regulator AGCOM (Italian AGCOM, 2018) also published guidelines prior to elections to ensure equal treatment of all political parties, political ad transparency and to encourage online fact-checking.

18. Japan Platform Services Study Group (2018-)

The Japanese Ministry of Internal Affairs and Communications (2018) formed a panel to study social media platforms' use of personal data in October 2018. Platform responses to disinformation are within the scope of the study as well, with references to the EU's Code of Practice. In December 2019, the Japanese news agency Jiji reported that social media platforms would be asked to "disclose their standards for removing fake news through artificial intelligence or other technologies and set up systems to deal with complaints". The final report of the study group was expected in early 2020 (Jiji, 2019).

19. Mexico National Electoral Institute (2018-2019)

The Mexican National Electoral Institute (INE) signed a cooperation agreement with Facebook, Twitter and Google to limit the spread of electoral disinformation and disseminate practical election information during their 2018 and 2019 elections. The social media companies also provide training to journalists and officials from the electoral office. During the elections, the INE also fact-checked information (Mexico INE, 2019). The president Córdova Vianello stressed that "[f]aced with the dilemma of regulating or limiting the use of social platforms and networks in political campaigns, the INE preferred to opt for a non-punitive model, to counter valid and true information against cases of false information, making an effort of public pedagogy and promoting citizen capacities

[493] https://www.bastabufale.it
[494] https://www.interno.gov.it/it/notizie/progetto-red-button-contro-fake-news

to distinguish qualities and sources of information" (translation by authors, Mexico INE, 2019).

20. Netherlands 'Stay Critical' Campaign (2019)

The Dutch Ministry of Interior Affairs ran a media literacy campaign, titled 'Stay Critical' (Dutch Government, 2019a) in the run up to the European Parliament elections in May 2019. In October 2019 the Dutch government adopted a strategy against disinformation, emphasising critical media literacy, transparency (preferably through self-regulation) of social media platforms and political parties, and maintenance of a pluriform landscape. In the strategy, the Dutch government also expressed explicit support for the EU Disinformation Code of Practice and EEAS East StratCom Task Force (Dutch Government, 2019b). Fact-checking is deemed important as a means of countering disinformation, but "addressing the content of disinformation as such is, according to the government, primarily not a task for governments or EU institutions, but primarily for journalism and science, whether or not in collaboration with internet services" (translation by authors, Dutch Government, 2019c).

21. New Zealand Parliamentary Inquiry into the 2017 General Election and 2016 Local Elections (2019)

The New Zealand Parliamentary Justice Committee published its report on the inquiry into their 2016 and 2017 elections in December 2019. Within the inquiry's broad scope, it raised concern about foreign interference during elections and made recommendations to prevent and minimise foreign influence through advertising, hacking, astroturfing, and disinformation.

The report also recommended the government consider the applicability of recommendations made on disinformation in the UK House of Commons' Digital, Culture, Media and Sport Committee's report on 'Disinformation and 'Fake News'' (which includes the creation of an independent regulator to monitor platform behaviour, improved media literacy, and improved transparency in online advertising) and the Australian Joint Standing Committee on Electoral Matters' Report on the 'Conduct of the 2016 Federal Election and Matters Related Thereto' (which includes the establishment of a permanent taskforce to prevent and combat cyber-manipulation, and greater clarity in regulatory framework on platforms) (New Zealand Parliament Justice Committee, 2019).

One week prior to the publication of the final report of the inquiry, a legislative act was introduced and passed in Parliament to ban foreign political donations during elections (*see* '*Adopted legislation*'), which could possibly have an impact on the disinformation landscape.

22. OAS Guide on Freedom of Expression regarding Deliberate Disinformation in Electoral Contexts (2019)

The Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights (IACHR), the Department of International Law (DIL) and the Department of Electoral Cooperation and Observation (DECO) of the Organization of American States developed a practical guide "to guarantee freedom of expression and access to information from various sources on the Internet in electoral contexts, without undue interference" (OAS, 2019).

They provide recommendations to a wide variety of actors in the American states: legislative branch, judiciary, executive branch, electoral authorities, Internet intermediaries, political parties, telecommunications companies, media and journalists, fact checkers, companies that trade data for advertising purposes, universities and research centres (OAS, 2019). Their ten pages of recommendations range from avoiding legislation that holds Internet intermediaries responsible for content produced by third parties and strengthening legislation that provides protections for citizens (data protection, freedom of expression, electoral advertising transparency) to promoting universal Internet access, media literacy campaigns, journalistic protection, and collaboration, transparency and due process in online content moderation (OAS, 2019).

23. South Africa Political Party Advert Repository (PAdRe) and Digital Disinformation Complaints Mechanism (Real411) (2019-)

The Electoral Commission of South Africa launched two initiatives in the run-up to the 2019 South African national and provincial elections to tackle digital disinformation and ensure citizens' access to information during the elections period. First, the Electoral Commission encouraged political parties to post their official advertisements in the Political Party Advert Repository (padre.org.za)[495]. Second, it launched a complaints portal for citizens to report digital disinformation during the elections (real411.org).[496] The Electoral Commission reported having received more than 70 complaints on digital disinformation within one month of launching the platform (van Diemen, 2019). The portal has since expanded its remit to also receive complaints on digital offences related to incitement to violence, hate speech and harassment of journalists as well. Both initiatives are supported by the civil society group Media Monitoring Africa (MMA).

24. Republic of Korea Disinformation Task Force (2018)

The ruling Democratic Party formed a task force to file complaints on grounds of defamation, incitement to violence and other problematic disinformation (Kajimoto, 2018; Sang-Hun, 2018).

25. Spain Hybrid Threats Unit (2019)

The Spanish government set up a 'hybrid threats' unit, consisting of different ministries, in the run-up to the 2019 EU and general elections. The aim of the unit was to have early response procedures to ensure election integrity with a focus on cyber security and monitor (and at times refuting) disinformation (Abellán, 2019).

26. Sweden Special Investigation into Development of Psychological Defence Authority (2018-)

In August 2018, the Swedish government mandated a special investigator to analyse and submit proposals to develop a Psychological Defence ('psykologiskt försvar') authority. The aim is to raise awareness and resilience to disinformation and hostile influence (Swedish government, 2018). In this line, the Swedish Civil Contingencies Agency (MSB, responsible for emergence and crisis management) describes their goal in psychological defence as "to ensure the Swedish people's willpower and resilience to pressure and disinformation from the opponent, within the framework of a - as far as possible under these extreme conditions - democratic, open society, with freedom of opinion and free

[495] https://padre.org.za
[496] https://www.real411.org

media. No censorship should occur" (translation by authors).[497] They provide media literacy tips on how to critique sources of information.

27. Ukraine 'Learn to Discern' Media Literacy Initiative (2015-)

The Ukrainian Ministry of Education partnered with the American NGO IREX to develop a media literacy programme to teach teenagers how to detect disinformation and hate speech. The programme has received support from UK and U.S. embassies in Ukraine (Ingber, 2019; Jankowicz, 2019).

In this context, the Ukrainian Ministry of Information Policy[498] was set up in 2014 and has the mandate to "counter information aggression", which amongst others has resulted in the blocking of some foreign websites (see '*Law enforcement and other state intervention*'), and also two proposed laws (although both were later retracted) which would have criminalised disinformation and also allowed the blocking of media outlets (Jankowicz, 2019).

28. UK House of Commons (Digital, Culture, Media and Sport Committee) Inquiry on Disinformation and 'Fake News' (2019)

The UK Commons Select Committee on Digital, Culture, Media and Sport made an 18-month inquiry into Facebook, Cambridge Analytica and AggregateIQ, published in February 2019. Its recommendations reflect those made by the Canadian Parliamentary Standing Committee and Australian government reports, highlighting the threat of disinformation to democracy and the need to enhance privacy and political advertising rules, electoral law. Specifically on social media regulation, the report recommended that "a compulsory Code of Ethics should be established, overseen by an independent regulator, setting out what constitutes harmful content", "a clear, legal liability for tech companies to act against agreed harmful and illegal content on their platform", and a 2% digital services tax on tech companies (UK Commons Select Committee on Digital, Culture, Media and Sport, 2019, pp. 89-90). The report further emphasised digital literacy, describing it as the "fourth pillar of education, alongside reading, writing and maths" (UK Commons Select Committee on Digital, Culture, Media and Sport, 2019, p.96).

In parallel, in 2017, the Information Commissioner's Office launched an investigation into misuse of personal data in political campaigns. Several organisations have been fined, including Facebook, who received the maximum £500,000 fine for data protection breaches (UK ICO, 2018). Finally, the UK Department for Digital, Culture, Media and Sport and the Home Office have also finalised an Online Harms White Paper in June 2019, proposing amongst others a 'duty of care' for social media platforms (*see 'legislative proposals*').

29. UK House of Commons Foreign Affairs Committee inquiry into Global Media Freedom (sub theme on disinformation) (2019)

This UK House of Commons Foreign Affairs Committee inquiry linked disinformation to threats to media freedom and journalists' safety. The Committee discussed "the problem of 'disinformation', how it threatens media freedom, and what role a free media can play

497 https://www.msb.se/psykologisktforsvar
498 http://mip.gov.ua/en/content/pro-ministerstvo.html

in response."[499] The final report from the inquiry was published in September 2019 (UK House of Commons Foreign Affairs Committee, 2019). It acknowledged that "[a] free media can also be an 'antidote' to the growing threat of 'disinformation'—the deliberate presentation of falsehoods as factual news, for personal, political, or commercial gain—while an unfree media risks being disinformation's mouthpiece." It also warned of how "... broad or overbearing laws were one of the most potent means of silencing journalists, with laws relating to defamation, national security or, increasingly, the 'illegalisation' of spreading 'fake news', disinformation, or rumours, being particularly open to abuse."

30. U.S. Senate Select Committee on Intelligence inquiry into "Russian Active Measures Campaigns and Interference In the 2016 US Election" (2017-2019)

The U.S. Senate Select Committee on Intelligence investigated the extent of foreign involvement in the 2016 election, publishing its report in several volumes (U.S. Senate Select Committee on Intelligence, (2019a) Volume 1, July 2019, (2019b) Volume 2, October 2019 and (2019c) Volume 3, February 2020). It found evidence of interference, primarily by a company called Internet Research Agency (IRA). It made recommendations to social media platforms to facilitate information sharing with government and law enforcement, to Congress on legislation for online political advertising transparency; and urged the Executive Branch to "establish an interagency task force to monitor foreign nations' use of social media platforms for democratic interference and develop a deterrence framework" (U.S. Senate Select Committee on Intelligence, 2019d).

Legislative proposals

A majority of recent legislative proposals (8 out of 13 analysed), aim to tackle disinformation through curation and the prism of intermediary liability obligations for online platforms regarding misinformation or hateful content. Similar to inquiries and task forces, the legislative proposals sometimes have an electoral-specific focus. Some other legislative proposals would criminalise the action of spreading disinformation. This can lead to a risk, highlighted on several occasions by fundamental rights activists, to be used against dissident journalists.

31. Argentina Bill on creating a Commission for the Verification of Fake News (2018)

A bill to create a Commision for the Verification of Fake News (Comisón de Verificación de Noticias Falsas, CVNF) within the National Electoral Chamber (Cámara Nacional Electoral, CNE) was making its way through the Argentine Parliament at the time of writing. Upon receiving a complaint, the Commission would fact-check disinformation during the national elections. Disinformation is defined as "[c]omments that appear to be news broadcast on the Internet or using other means, usually created to influence political opinion or as a joke, that has no correlation with the reality of the facts," although "opinion columns expressed in ways that are based on ideological positions or reasoning expressed discursively" are exempt from scrutiny (unofficial translation, Argentina Bill on creating a Commission for the Verification of Fake News, 2018b).

[499] https://www.parliament.uk/business/committees/committees-a-z/commons-select/foreign-affairs-committee/news-parliament-2017/global-media-freedom-evidence-17-19/

In a case of disinformation, the Commission would notify the National Electoral Chamber, who could request social media platforms to mark the post with a "notice of doubtful credibility" (unofficial translation), decrease its circulation, and create a database of disinformation freely available to the public. Proposed sanctions include warnings, fines, disqualification as a government contractor/provider for up to ten years, loss of benefits or special tax regimes, and suspension for two years of platform that failed to deprioritise content after being ordered to do so (Argentina Bill on creating a Commission for the Verification of Fake News, 2018b).

32. Chile Proposal to End Mandate of Elected Politicians due to Disinformation (2019)

In January 2019, five Chilean senators proposed a law to end the mandate of elected politicians if established that they disseminated, promoted or financed disinformation on opponents during election campaigns. The bill aims to curtail the use of disinformation for political gain. The proposal was sent to the Senate's Constitution, Legislation, Justice and Regulation Commission, but has not yet been analysed (Chile Senate, 2019).

33. France Fight against Online Hate Speech Law Proposal (2019)

In 2019, the French Parliament started deliberating a proposed law to fight "incitement to hatred and insults on grounds of race, religion, ethnicity, sex, sexual orientation or disability" (translation by authors) published online. The law proposal as submitted by Member of Parliament in March 2019 stipulates that online platforms with numerous connections in the French territory need to remove or block access to manifestly illegal content within 24 hours of receiving complaints made by users. Failure to comply by online platforms could result in fines. Under the proposed law, online platforms would also need to provide mechanisms for flagging content and appealing decisions, as well as providing transparency reports on their actions. The proposal mandates that the French Broadcasting Authority (CSA) monitor platform behaviour. The legislation would build on the 1881 freedom of the press law (French Parliament, 2019; EurActiv, 2019).

34. German Network Enforcement Act Update (2020)

In February 2020, the German government approved a regulatory package that would complement the 2017 Network Enforcement Act (NetzDG, *see* '*adopted legislation*') and would require platforms to report illegal hate speech to the police and to provide the users' IP addresses. At the time of writing, the bill still needed to be approved by the parliament. A second text revising the initial NetzDG was expected to be on the table in mid 2020, focusing on the complaint management of the platforms (German BMJV, 2020a; German BMJV, 2020b).

35. India Proposed Amendments to IT Intermediary Guidelines (2018)

The Indian Ministry of Electronics and Information Technology released Draft Information Technology Intermediaries Guidelines (Amendment) Rules for public comments in December 2018. Under section 3 on due diligence, a first amendment proposed that upon court order or notification of the appropriate government agency, intermediaries have a 24 hour window to remove or disable access to content deemed unlawful - "relatable to Article 19(2) of the Constitution of India such as in the interests of the sovereignty and integrity of India, the security of the State, friendly relations with foreign States, public order, decency or morality, or in relation to contempt of court, defamation or incitement to an offence". A second amendment called on intermediaries to "deploy

technology-based automated tools or appropriate mechanisms, with appropriate controls, for proactively identifying and removing or disabling public access to unlawful information or content" (Indian MeitY, 2018). After a consultation period, the proposal was discussed in inter-ministerial format in July 2019 and an amended text is expected in early 2020 (Agrawal, 2019).

36. Ireland Proposal to Regulate Transparency of Online Political Advertising (2019)

In November 2019, the Irish government introduced a legislative proposal to regulate the transparency of online paid political advertising within election periods, in line with a recommendation of the Irish Interdepartmental Group on Security of the Election Process and Disinformation (*see* '*Inquiries, task forces and guidelines*', Irish Department of the Taoiseach, 2019).

37. Israel Proposed Electoral Law Amendments and 'Facebook Laws' (2018)

The Israeli Knesset (Parliament) introduced and considered multiple bills to amend the electoral law on transparency regarding election advertising and countering foreign propaganda during elections. Several bills pertaining to the removal of illegal content online were also before the Parliament (Levush, 2019) but did not make it through the Knesset prior to the April 2019 elections.

Prior to the April 2019 elections, the Israeli Central Elections Committee chairman and Supreme Court Justice met with Facebook to request a ban of anonymous elections adverts, to which the company complied (Staff, 2019; Milano, 2019). This action follows one of the main recommendations of the Committee for Examination of the Elections (Modes of Propaganda) Law to "extend the application of substantive provisions of the Law to the internet and to social platforms" (Levush, 2019).

38. Nigeria Protection from Internet Falsehood and Manipulation Bill (2019)

The Nigerian Protection from Internet Falsehood and Manipulation Bill closely resembles the Singapore Protection from Online Falsehoods and Manipulation Act.

The bill proposes the prohibition of the transmission of false statements of fact, the making or altering bots for communication of false statements of fact, and providing services for transmission of false statements of fact in Nigeria. The bill broadly targets false statements of fact that are prejudicial to the security of Nigeria; public health, safety, tranquility, finances; friendly relations with other countries; influence the outcome of an election; incite feelings of enmity, hatred or ill-will between different groups of people; and diminish public confidence in government.

Sanctions for spreading a false declaration of fact include fines and up to three years of imprisonment. When directed, individuals would need to post a correction notice and stop the transmission of the false declaration. The bill also proposes the possibility for the Nigerian Law Enforcement Department (via the Nigerian Communications Commission) to order Internet access providers (who in case of non-compliance can be fined) to block access to a declared online location.

The law also includes the possibility to appeal to the Law Enforcement Department and then the courts (Nigeria Protection from Internet Falsehood and Manipulation Bill, Nigerian Senate, 2019; Paquette, 2019).

39. Philippines Anti-False Content Bill (2019)

An Anti-False Content Bill was introduced in the Filipino Senate in July 2019. It has been referred to the Committee on Public Information and Mass Media for consideration. The bill prohibits "publishing information that the publisher knows or reasonably believes to be false or misleading to the public," using a fictitious account in publishing such false content, providing a service and financing that activity, or failing to comply with orders in Section 5 of the Act. Section 5 of the Act permits the Department of Justice Office of Cybercrime to make orders for rectification, take down, or removal of the published information. Sanctions include fines and up to twelve years of imprisonment. The bill allows for appeal to the Office of Secretary of the Department of Justice. (Philippines Senate, 2019 Bill no 9, Anti-False Content Bill).

40. Republic of Korea Proposed Law (2018)

The Republic of KoreaNational Assembly is considering law amendments, ranging from increasing the responsibility of Internet intermediaries to manage disinformation to extending defamation laws to criminalise disinformation (Sang-Hun, 2018; Corcoran, Crowley & Davis, 2019).

41. Sri Lanka Proposed Amendments to Penal Code (2019)

After the April 2019 terrorist attack on churches and hotels on Easter Sunday social media was blocked in the days following the attack in order to limit incitement to violence against muslims (*see* '*Law enforcement and other state intervention*'). In May 2019, the Sri Lankan cabinet approved amendments to the penal and criminal procedure codes to prohibit fake news and hate speech that is "harmful to harmony between the nations and national security". Sanctions include fines and up to five years imprisonment (AFP, 2019a; Sri Lanka Brief, 2019).

42. UK Online Harms White Paper (2019)

The UK government ran a consultation on an 'Online Harms White Paper' in 2019 which included limited hearings.[500] In February 2020, the government gave its response. Even though this response is not in itself a regulatory proposal, it lays clear ground for further governmental regulatory action. It foresees a statutory duty of care for social media platforms to protect users from harmful and illegal terrorist and child abuse content. Based on the feedback from the consultation, the government highlighted that "to ensure protections for freedom of expression, regulation will establish differentiated expectations on companies for illegal content and activity, versus conduct that is not illegal but has the potential to cause harm", which might include disinformation. The government response to the White Paper envisages that an independent regulator (possibly Ofcom) would monitor platform behaviour.[501]

500 https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

501 https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

43. U.S. Tennessee State Legislature bill to register CNN and The Washington Post as "fake news" agents of the Democratic Party (2020)

In February 2020, the Tennessee Legislature (2020) debated House Joint Resolution 779, brought by a State Representative (Republican) "to recognize CNN and The Washington Post as fake news and condemn them for denigrating our citizens." The Tennessee General Assembly Fiscal Review Committee (2020) summarised the Bill thus: "[r]esolves to recognize CNN and The Washington Post as fake news and part of the media wing of the Democratic Party, and further resolves to condemn such media outlets for denigrating our citizens and implying that they are weak-minded followers instead of people exercising their rights that our veterans paid for with their blood." Following debate, it was recommended for passage by the House Judiciary Committee in a 13-5 vote in February 2020. At the time of writing, it was proceeding for further legislative review.

Adopted legislation

We identified 28 countries that have passed legislation related to disinformation. Governments choose to tackle the issue of disinformation, either by updating existing regulations or with passing new legislation. The scope of the established legislation varies from media and electoral laws to cybersecurity and penal codes. The regulations either target the perpetrators (individuals and media) of what the authorities deem as being disinformation, or shift the responsibility to the Internet communication companies, with obligations to moderate and eventually remove content. In laws where disinformation is defined broadly or provisions included in penal codes, there is a major risk of disinformation being used as a means to gain control over citizens and journalists, and thus a risk of censorship if no independent control is foreseen.

44. Argentina Political Party Financing Law (2019)

Several amendments to the Argentine Electoral Law pertaining to transparency of political party financing were passed in May 2019, prior to general elections held in October 2019. Among others, the legislative changes give the Argentine National Electoral Chamber the mandate to maintain a registry of social media accounts and websites of candidates and political parties, and a registry of survey and opinion polling companies. Candidates and political parties are also obliged to disclose their online campaign finances, and additional public financing for journalistic reporting is guaranteed (Argentina Political Party Financing Law, 2019).

These amendments follow through on an extraordinary agreement reached in August 2018 in the National Electoral Chamber, in which the ambition of the creation of a registry of political candidates was to aid the detection of false accounts (Rodriguez-Ferrand, 2019).

45. Bangladesh Digital Security Act (2018)

In October 2018, Bangladesh (2018) passed a Digital Security Act, replacing parts of its 2006 Information Communication Technology Act and strengthening its 1923 Official Secrets Act.

The law allows searches and arrests without court warrant, imprisonment up to fourteen years, and fines for cyber-related crimes, such as cyber-terrorism, identity fraud, hacking and illegal access and distribution of state information.

The law also includes provisions on online speech regulation, such as spreading "propaganda or campaign" against Bangladesh's 1971 war for independence from Pakistan, "offensive, false or fear inducing data-information", information that "hampers the religious sentiment or values" and defamation of individuals.

The law has been used to request content take downs and arrests related to the spreading of rumours and false information on online platforms (Quadir, 2018; Alam, 2018).

46. Belarus Media Law (revised in 2018)[502]

In June 2018, the Belarusian Parliament passed amendments to their Media Law to be able to prosecute individuals and block social media and websites for spreading disinformation online (RFE/RL, 2018a). The legislation includes a ban on foreign media outlets, tighter regulations of media registration and journalist accreditation, and a responsibility for online publishers to monitor and prevent defamation and disinformation (EFJ, 2018).

47. Benin Digital Code (2017)

Chapter 4 of the Benin Digital Code, passed in 2017, deals with abusive content and online press offenses, covering crimes such as incitement to hatred, racism, violence, terrorism, denial of crimes against humanity and defamation. Specifically, Article 550 of the Benin Digital Code tackles online harassment and prohibits spreading disinformation about a person. Sanctions for spreading disinformation include fines and imprisonment up to six months (Benin Digital Code, 2017).

48. Brazil Law Criminalising Electoral Disinformation (2019)

In the run up to the 2018 general elections, several proposed laws were introduced to criminalise electoral disinformation. In September 2019, the Brazilian Congress passed a law amending the electoral code, which defines the crime of "slanderous denunciation for electoral purpose", with a penalty of two to eight years of imprisonment. Congress has also formed a joint parliamentary inquiry committee (CPI das Fake News) to investigate use of profiling and disinformation to influence the outcome of the 2018 elections (Redação Exame, 2019).

49. Burkina Faso Penal Code (revised in 2019)

Amendments made to the penal code in Burkina Faso in June 2019 criminalise 'demoralising' defence and security forces, providing false information about a destruction of property or an attack on people, information compromising defence and security forces, information depicting terrorism, and insulting the memory of a deceased person. Sanctions include fines and imprisonment up to ten years. A judge can mandate blocking access to websites or content distributing false information (RSF, 2019; Burkina Faso Constitutional Council, 2019).

[502] When the phrase 'revised in [year]' is used as a reference, this means the amendments made to the relevant law were specific to disinformation.

50. Cambodia Fake News Directives (2018)

In the run up to the 2018 elections, the Cambodian government passed two directives to allow for blocking of websites on account of security threats and to criminalise the posting of disinformation online. Sanctions include fines and imprisonment up to two years. Media outlets (offline, but also if ending on .kh) are obliged to register with the Cambodian Ministry of Information (Lamb, 2018; Sovuthy, 2018). The ministry has threatened to revoke media licenses if they spread disinformation that endangers national security. Offending content has also been removed (Livsier, 2019).

51. Cameroon Penal Code (1967, revised) and Cyber Security and Cyber Criminality Law (2010)

Article 113 of the Cameroonian penal code prohibits "sending out or propagating false information liable to injure public authorities or national unity" (Cameroon Penal code, 1967, revised). Sanctions include fines and imprisonment up to three years. Disinformation has also been tackled through other provisions, such as on defamation and secession (CPJ, 2019a; Funke, 2019). Article 78 of the 2010 law on cybersecurity and cyber criminality sanctions publishing or propagating "news without being able to prove its veracity or justifying why there are good reasons to believe the veracity of the news" through electronic means with fines and imprisonment up to two years. The sanctions are doubled if "the offence is committed with the aim of undermining public peace" (translation by authors, Cameroon Cyber Security and Cyber Criminality Law, 2010).

52. Canada Elections Modernisation Act (2018)

In the run-up to the 2019 federal election, the Canadian Parliament passed the 'Elections Modernization Act' (2018). The law increases electoral advertising transparency of political parties and third parties. It also requires social media companies to increase transparency of online political advertising. As a result, Canada has obliged social media companies to "maintain a registry of partisan and election advertising published during the pre-election and election periods".[503]

53. China Anti-Rumour Laws (2016, 2017, 2018)

China has updated and passed several laws to tackle disinformation or "rumours". For instance, the Cybersecurity Law (2016) criminalises the deliberate spreading of false information and rumours that undermine the economic and social order (Repnikova, 2018). The law also deals more broadly with data management and network security. Similar provisions can be found in the Criminal Law (updated 2017), where subversion of public order and slander are addressed. Sanctions can include imprisonment, criminal detention, control and deprivation of political rights (People's Daily 2013; PKULaw, 2017; Repnikova, 2018).

Further, pertaining to regulating Internet intermediaries, the Administrative Regulations on Internet News Information Services (2017) indicate that online platforms and news services need to repost and link to official government-approved sources, while the Administrative Regulations on Microblog Information Services (2018) oblige microblogs (e.g. Weibo) to verify a blogger's identity and have 'anti-rumour' mechanisms in place to proactively monitor, block and refute disinformation (Yang 2018; Repnikova, 2018).

[503] https://www.canada.ca/en/democratic-institutions/news/2019/01/encouraging-social-media-platforms-to-act.html

54. Côte d'Ivoire Penal Code (1981, revised)[504] and Press Law (revised in 2017)

Article 173 of the Penal Code in Côte d'Ivoire (Côte d'Ivoire Penal Code 1981, revised, Article 173) prohibits "the publication, dissemination, disclosure or reproduction by any means whatsoever of false news, material that is fabricated, falsified or deceptively attributed to third parties" (translation by authors). Sanctions include fines and up to three years imprisonment. Article 97 of the Law on the Legal Regime for the Press (Côte d'Ivoire Penal Code 1981, revised. Article 97, 2017) was extended to include online press and similarly prohibits "the publication, dissemination, disclosure or reproduction through the press of false news, material that is fabricated, falsified or deceptively attributed to third parties" and imposes fines (translation by authors) (Drissa, 2019).

55. Egypt Anti-Fake News Laws (2018)

Egypt has several laws that relate to freedom of expression and disinformation. First, Law No. 180 of 2018 on Regulating the Press and Media allows the Supreme Council for Media Regulation to suspend, ban and block media outlets and websites (with more than 5000 followers) that pose a threat to the national security, disturb the public peace, promote discrimination, violence, racism, hatred, or intolerance, and spread disinformation. The law also imposes additional administrative and licensing requirements on media outlets (Sadek, 2019; TIMEP, 2019). Second, Law No. 175 of 2018 on Anti-Cybercrime stipulates that an investigating authority can block or suspend websites if it poses a threat to the national security or national economy, or contains other content criminalised under the Anti-Cybercrime Law. Sanctions also include fines and imprisonment up to two years. The law also includes provisions on criminal liability of service providers (Sadek, 2019; Bälz & Mujally, 2019; Magdy, 2019). Finally, Article 80(d) of the Penal Code (1937, revised) provides the possibility to impose fines and up to five years of imprisonment on "whoever deliberately spreads false information or rumors abroad about the internal conditions of the country that might weaken the country's financial credibility or harm the country's national interests" (Sadek, 2019).

56. Ethiopia False Information Law (2020)

In February 2020, Ethiopia's parliament passed a law to fight hate speech and the dissemination of false information online. Dissemination of such content on platforms with more than 5000 followers can lead to fines or imprisonment of various lengths, up to five years (Endeshaw, 2020).

57. France Fight against Manipulation of Information Law (2018)

In 2018, the French Parliament passed a law about the manipulation of information before and during election periods ("Anti-Fake News Law"). The legislation builds on the 1881 freedom of the press law. French senators appealed the law on grounds that it would have restricted freedom of expression. However, the law was validated by the Constitutional Council and enacted in December 2018 (Conseil Constitutionnel, 2018)

The law defines disinformation as "inaccurate or misleading accusations or allegations with the aim of changing the sincerity of a vote". Three months prior to an election, upon a complaint from public authorities, electoral candidates, political groups and individuals, an interim judge is authorised to act "with appropriate and necessary measures" to halt

[504] When we use the phrase '[year], revised' as a reference, this means the relevant law was amended (often multiple times), but that the latest changes to the law were *not* specific to disinformation.

the dissemination of disinformation. The judgement must follow within 48 hours of receiving a complaint. The law also places a "duty of cooperation" on online platforms to provide mechanisms for flagging content, algorithmic transparency, promotion of mainstream news content, advertising transparency, and media literacy initiatives. Finally, the French law grants additional power to the French Broadcasting Authority (CSA) to monitor platform behaviour and revoke license of foreign run broadcasters spreading misinformation (France Fight against Manipulation of Information Law, 2018; Damiano Ricci, 2018).

58. Germany Network Enforcement Act (2017)

The German Parliament adopted the Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) in June 2017. The law obliges for-profit social media platforms with more than two million registered users in Germany to take action against hate speech and offences as described in the German criminal code (such as dissemination of propaganda and use of symbols of unconstitutional organisations, incitement to violence, crime and terrorism, religious insults, defamation, and distribution, acquisition and possession of child pornography).

Social media platforms are obliged to have clear procedures for flagging content and handling complaints, to remove or block access to "manifestly unlawful" content within 24 hours, "unlawful" content within 7 days, and in case of having received more than 100 complaints per calendar year about unlawful content, to publish reports every six months on how they dealt with flagged content. The person submitting the complaint, as well as the affected user are notified and provided reasons for the decision. (German NetzDG 2017; German NetzDG English translation 2017).

59. Indonesia Electronic Information and Transactions Law (revised in 2016)

In October 2016, Indonesia revised its 2008 Electronic Information and Transactions Law. The law seeks to regulate online content and authorises the government to order Electronic System Providers to terminate access to content deemed to be in violation, such as pornography, terrorism, insult or defamation, hate speech, but also other content deemed "negative". Sanctions include fines and up to six years of imprisonment. The law also includes a right to be forgotten (data deletion based on a court ruling) (Molina, 2016; The Jakarta Post, 2016).

60. Kazakhstan Penal Code (2014)

Article 274 of the Kazakh Penal Code prohibits "the dissemination of knowingly false information, creating a danger of violation of public order or infliction of substantial harm to the rights and legal interests of citizens or organisations, or the interests of society or the state, protected by the Law" (unofficial translation, Kazakhstan Penal Code, 2014). There is a scale of sanctions, including fines, correctional works and up to seven years of restrictions on freedom, or imprisonment (up to ten years if committed by a criminal group or in wartime or emergency situations). Article 130 of the Penal Code on defamation can lead to fines, correctional works, and up to three years of restrictions on freedom or imprisonment.

61. Kenya Computer Misuse and Cybercrimes Act (2018)

The Kenyan Computer Misuse and Cybercrimes Act was passed in May 2018 and encompasses a broad range of computer crimes, from unauthorised access, interception and interference, cyber espionage and cyber terrorism to child pornography, cyber harassment, computer fraud, phishing and identity theft.

Sections 22 and 23 of the Act pertain to false publications and target "[a] person who intentionally publishes false, misleading or fictitious data or misinforms with intent that the data shall be considered or acted upon as authentic, with or without any financial gain". Sanctions for publishing such disinformation can include a fine and up to two years imprisonment. If the publication of the false information "is calculated or results in panic, chaos, or violence among citizens of the Republic, or which is likely to discredit the reputation of a person commits an offence and shall on conviction," the imprisonment can increase up to ten years (Kenya Computer Misuse and Cybercrimes Act, 2018).

The constitutionality of the Act has been challenged in court and a judgment is expected in early 2020. Meanwhile implementation of 26 sections of the Act, including those on false information, has been suspended (Itimu, 2019).

62. Malaysia Anti-Fake News (Repeal) Act (2019)

Malaysia passed an Anti-Fake News Act in April 2018. It defined 'fake news' broadly ("any news, information, data and reports, which is or are wholly or partly false, whether in the form of features, visuals or audio recordings or any other form capable of suggesting words or ideas"), and provided for sanctions of fines or up to six years imprisonment. The Malaysian Parliament sought to repeal the Act twice: the first attempt was blocked by the Senate in August 2018, but was successful in December 2019 (Free Malaysia Today, 2019).

After a change in government, the Anti-Fake News Act was deemed unnecessary as existing laws (Penal Code, Sedition Act, Printing Presses and Publications Act, Communications and Multimedia Act) already tackle disinformation. Under these sets of legislation, individuals can face fines and imprisonment up to seven years for a broad set of actions pertaining to disinformation, such as speech that is "prejudicial to public order, morality, security, or which is likely to alarm public opinion, or which is or is likely to be prejudicial to public interest or national interest", defamation, sedition, hate speech, and incitement to violence (Buchanan, 2019).

63. Myanmar Telecommunications Law (2013) and Penal Code (1861, revised)

Section 66(d) of the Telecommunications Law prohibits "extorting, coercing, restraining wrongfully, defaming, disturbing, causing undue influence or threatening to any person by using any Telecommunications Network". Sanctions include fines and up to three years imprisonment (Myanmar Telecommunications Law, 2013).

Section 505(b) of the Penal Code has also been used to curb what the authorities deemed to be disinformation. It criminalises "any statement, rumour or report" that is "likely to cause, fear or alarm to the public or to any section of the public whereby any person may be induced to commit an offence against the State or against the public tranquility". Sanctions include fines and up to two years imprisonment (Myanmar Penal Code, 1861, revised).

64. New Zealand Electoral Amendment Act (2019)

Although the act does not directly address disinformation, it may have resonance. It was introduced in December 2019 amidst concerns about foreign interference (including hacking and disinformation campaigns) during New Zealand elections. The act bans foreign political donations over NZ$50 and requires transparency in election advertising on all mediums (Ainge Roy, 2019).[505]

65. Oman Penal Code (1974, revised)

The Omani Penal Code contains provisions on knowingly spreading "false news on crimes that have not been committed" and "rumours that affect the state". Sanctions include fines and up to three years of imprisonment. (Al Busaidi, 2019; Kutty, 2018).

66. Pakistan Prevention of Electronic Crimes Act (2016)

The Pakistani Prevention of Electronic Crimes Act (2016) criminalises actions such as glorification of terrorism, cyber-terrorism (which includes threats to the government and hate speech), and offences against the dignity or modesty of a person (such as false information spread to intimidate or harm an individual). Sanctions include fines and imprisonment. The Act also allows for removal or blocking of information. The Ministry of Information and Broadcasting refers to the 2016 Act in raising awareness and refuting disinformation through their @FakeNews_Buster[506] Twitter handle.

67. Philippines Penal Code (revised in 2017)

Article 154 of the Filipino Penal Code was amended in 2017 to prohibit publishing false news "which may endanger the public order, or cause damage to the interest or credit of the State". Sanctions include fines and imprisonment up to six months (Philippines Revised Penal Code, Act no 10951, 2017).

68. The Russian Federation's Fake News Amendments to Information Law and Code on Administrative Violations (2019)

In March 2019, the Russia Federation adopted two laws, amending the Federal Law on Information, Information Technologies and the Protection of Information (Information Law) and the Code on Administrative Violations.

The amendments to the Information Law build on changes made in 2016, which established a liability regime for the Russian Federation's news aggregators (with more than 1 million daily users) and provided the Russian Federal Service for Supervision of Communications, Information Technology and Mass Media (Roskomnadzor) with monitoring and blocking powers. Both state bodies and courts can request blocking of content via the news aggregators (Grigoryan, 2019; Richter, 2019). The 2019 amendments to the Information Law prohibit the online dissemination of "unreliable socially significant information", which would constitute a threat to citizens, property, public order and/or public security, transportation, industry or communication (Richter, 2019). They also allow the Russian Federation's media regulator Roskomnadzor upon request of the Prosecutor-General and his deputies to order the removal of disinformation online, and if necessary

505 https://www.parliament.nz/en/pb/bills-and-laws/bills-proposed-laws/document/BILL_93304/electoral-amendment-bill-no-2

506 https://twitter.com/FakeNews_Buster

the blocking of infringing websites (Grigoryan, 2019; Richter, 2019). The law includes the possibility to appeal in court (Richter, 2019).

The amendments to the Code on Administrative Violations in turn set the fines for spreading disinformation (Grigoryan, 2019; Richter, 2019). Finally, it is worth noting that the Russian Federation's Criminal Code also tackles disinformation, prohibiting for instance denial of Nazi crimes and falsification of history. Sanctions here include fines, compulsory community service and imprisonment (Grigoryan, 2019; Richter, 2019).

69. Singapore Protection from Online Falsehoods and Manipulation Act (2019)

The Singaporean Parliament passed the Protection from Online Falsehoods and Manipulation Act (POFMA) in May 2019. The law came into effect in October 2019, with temporary exemptions being granted to certain Internet intermediaries to comply (See Kit, 2019).

POFMA's wording identifies a scope of targeting those false statements of fact that are perceived to be prejudicial to the security of Singapore; public health, safety, tranquility, finances; friendly relations with other countries; influence the outcome of an election; incite feelings of enmity, hatred or ill-will between different groups of people; and diminish public confidence in government.

In this scope, the law prohibits the online communication of false statements of fact, the making or altering of bots for communication of false statements of fact, and providing services for communication of false statements of fact in Singapore.

Communications from the Singapore authorities for this report state that: "POFMA provides for corrections as its primary tool, which would require a notice to be placed alongside the original online post or article stating that it contains a false statement of fact and directing viewers to a Government website for the facts of the case. It serves as a 'right of reply' mechanism, without the original post being removed, and the intent is to allow readers to decide for themselves the truth of the matter."

The communications add: "If the correction directions are not complied with, the Singapore Government may order internet access service providers to block access to a specified online location by end-users in Singapore. Non-compliance on the part of the internet access service provider can result in a fine. In extreme cases where there is a threat of serious harm, the Singapore Government may issue directions to individuals or internet intermediaries to make online falsehoods unavailable to Singapore viewers." Further the communications state: "Beyond providing access to and increasing the visibility of correction notices, measures under POFMA include disrupting inauthentic accounts that amplify falsehoods; discrediting online sources of falsehoods; and cutting off financial incentives to online sources of falsehoods."

Sanctions for spreading online falsehoods can apply both to individuals and organisations. They can include fines and up to ten years of imprisonment. When directed by the Singaporean government, individuals or internet intermediaries need to post a correction notice and stop the communication of false information. According to the Singapore authorities' communications received by the UNESCO Secretariat, "If a POFMA direction is not complied with, an access blocking order can be made to internet access service providers but not Internet Intermediaries. However, Internet Intermediaries can be

directed to block access to content on their platform if the owner or operator of a page on the platform did not comply with the requirements of a 'declared online location' (but this does not apply for non-compliance to a POFMA direction)."

The law also includes the possibility to appeal in court. (Singapore POFMA, 2019)

70. Thailand Computer Crime Act (revised in 2017) and Cybersecurity Act (2019)

The Thai Computer Crime Act tackles crimes, such as illegal access and damage to a computer system, illegal interception and disclosure of information, as well as online defamation. Specific to disinformation, Section 14(2) prohibits "bringing into a computer system computer data which is false, in such a manner likely to cause damage to the maintenance of national security, public safety, national economic security, or infrastructure for the common good of the Nation, or to cause panic amongst the public" (Thailand Computer Crime Act, 2017; Chitranukroh, 2017).

The 2017 amendments also provide enforcement authorities with the possibility to access computers (with a court warrant) and for a Computer Data Screening Committee (appointed by the Digital Economy and Society Ministry) to issue a stop to the dissemination of illegal computer data (Thailand Computer Crime Act, 2017; Chitranukroh, 2017). Sanctions include fines and up to five years of imprisonment. The 2019 Cybersecurity Act expands government reach online and additionally allows it "to summon individuals for questioning and enter private property without court orders in case of actual or anticipated 'serious cyber threats'" (Tanakasempipat, 2019a).

71. Vietnam Cyber Security Law (2018)

Vietnam passed a Cyber Security Law (CSL) in 2018. The law prohibits a wide range of acts, such as breaching existing laws on national security, public safety and order, building opposition to the State of Vietnam, incitement to violence, "publishing information which is lewd, depraved or criminal", cyber espionage, terrorism and hacking.

Specific to disinformation, the law prescribes sanctions against "providing false information, causing confusion among the citizens, causing harm to socioeconomic activities, causing difficulties for the operation of State agencies or of people performing public duties, or infringing the lawful rights and interests of other agencies, organizations and individuals" (Bich Ngoc, 2019; Chung Seck & Son Dang, 2019).

Further, the Cyber Security Law establishes a data localisation requirement on Internet intermediaries, in order to provide user information for investigations. Intermediaries are also obliged to remove or block access to unlawful content and accounts, both proactively and within 24 hours of receiving a request from a specialised force in charge of cybersecurity protection under the Ministry of Public Security or competent authorities under the Ministry of Information and Communications (Bich Ngoc, 2019; Chung Seck & Son Dang, 2019).

Law enforcement and other state intervention

This section highlights examples of the enforcement of existing regulations or laws that are used to address what is deemed to be disinformation. These actions include fines, arrests, internet shutdowns and website takedowns. Enforcement has been applied to: individuals (including journalists and activists); news organisations; foreign state media considered to be disseminating disinformation; or the internet communication companies judged to be responsible for the massive reach of disinformation. Internet shutdowns also have been applied to prevent the spread of disinformation in electoral periods, which may be seen as blunt (over/under-inclusive) measures in limiting access to information. This content is presented without judgment by UNESCO or the Broadband Commission on any decisions taken under sovereign legislation, or on the comprehensiveness of the cited source(s).

72. Bahrain

Accounts reported of arrests, fines and prison sentences on grounds of defamation and spreading false information that jeopardise the security of the state (BBC, 2018a; Associated Press, 2019).

73. Bangladesh

In the days preceding the general elections on 30 December 2018, shutdown of mobile Internet services "to prevent rumours and propaganda surrounding the vote", a spokesman for the Bangladesh Telecommunication Regulatory Commission reportedly said (Al Jazeera, 2018; Paul, 2018).

74. Benin

Article 550 of the Digital Code (*see* '*Adopted legislation*') has been applied to prosecute, fine and imprison journalists for spreading disinformation online (Houngbadji, 2020; Fitzgibbon, 2020).

75. Cambodia

At least one Cambodian citizen reportedly convicted and jailed for violating the Anti-Fake News Law (*see* '*Adopted legislation*'), and in particular for spreading disinformation that threatens national security (Kongkea, 2019a & Kongkea, 2019b).

76. Cameroon

Journalists jailed on charges of spreading false news that jeopardises the security of the state, on the basis of the Cameroonian penal code and cyber criminality law (*see* '*Adopted legislation*'). The accused have had to appear before military court as well (CPJ, 2019a; Funke, 2019).

77. People's Republic of China

Through the Anti-Rumour Laws (*see* '*Adopted legislation*') and other relevant legislation, the government, in collaboration with internet intermediaries, actively blocks content, shuts down accounts and prosecutes individuals for spreading what has been deemed to be disinformation and rumours that undermine the social, economic and public order (Qiu & Woo, 2018; Repnikova, 2018).

78. Côte d'Ivoire

In 2019 a politician in Côte d'Ivoire was fined and sentenced to one year imprisonment for a tweet that "spread fake news that incited violence" (AFP, 2019c; BBC, 2019a).

79. Egypt

Various regulations, including their 2018 law Regulating the Press and Media and Anti-Cybercrime Laws (*see* '*Adopted legislation*'), have been enforced to detain and fine individuals, and block websites for publishing information that authorities have deemed to have threatened national security and spread disinformation (BBC, 2018b; Magdy, 2019).

80. Germany

In the context of the German Network Enforcement Act (*see* '*Adopted legislation*'), Facebook was fined 2 million Euro by the German Federal Office of Justice in 2019 for lack of transparency in its reporting on the complaints filed and actions taken when tackling hate speech and other criminal offences (Prager, 2019; Zeit, 2019).

81. India

Internet access in India has been regularly shut down, sometimes for extended periods (Gettleman, Goel & Abi-Habib, 2019; Burgess, 2018).The Temporary Suspension of Telecom Services Rules, passed in 2017, allows authorities to regulate "the temporary suspension of telecom services due to public emergency or public safety" (Indian Ministry of Communications, 2017).

82. Indonesia

In collaboration with internet communications companies, "close to a million websites" (Board, 2019) have been blacklisted under the Electronic Information and Transactions Law (*see* '*Adopted legislation*') and other relevant legislation. Arrests have also been made for spreading information that violates Indonesian laws (Tapsell, 2019), and access to social media has reportedly been restricted to prevent 'hoaxes', imposing limits on the ability to upload videos or photos (Beo Da Costa, 2019).

83. Kazakhstan

In 2017, Forbes Kazakhstan and ratel.kz were found guilty of defamation in a lawsuit. The news outlets were fined and ordered to remove the defamatory content and issue a retraction notice. Forbes Kazakhstan complied. Ratel.kz paid the fine, but requested clarification from the court on the content to be removed. Not having received a response, ratel.kz did not remove the content, nor issue a retraction. In 2018, the court ordered a one-year shutdown of ratel.kz for violating rules for registration, use and distribution of domain names in Kazakhstan. In both instances, the investigations were opened against the news outlets for "disseminating knowingly false information" (Human Rights Watch, 2018a; RFE/RL, 2018b; Keller, 2019).

84. Latvia

In 2016, the Latvian Network Information Center ordered the shutdown of the local website of the Russia Federation's foreign news channel Sputnik deemed to be "a propaganda tool", after the Foreign Affairs Ministry drew attention to Sputnik's coverage of Ukraine and routine denial of the country's territorial integrity (Latvian Public Broadcasting, 2016; EurActiv, 2016).

85. Malaysia

Malaysia has enforced various relevant legislation (including the now repealed Anti-Fake News Act) to prosecute individuals for spreading disinformation. As an example, in January 2020, the Malaysian Communications and Multimedia Commission (2020) detained four individuals suspected of spreading false news on the coronavirus under Section 233 of the Communications and Multimedia Act.

86. Myanmar

The Burmese Telecommunications Law, Penal Code (*see* '*Adopted legislation*') and other relevant legislation have been used to curb content deemed by the authorities to be disinformational in Myanmar (Associated Press, 2018; Schulman, 2019).

87. The Russian Federation

Enforcing the Information, Information Technologies and the Protection of Information Law (*see* '*Adopted legislation*') and other relevant laws (BBC, 2019b; Richter, 2019), the Russian Federation's media regulator (Roskomnadzor) blocks websites and content deemed to disrespect the Russian Federation's authorities (Zharov, 2019).

88. Singapore

Based on the Protection from Online Falsehoods and Manipulation Act (*see* '*Adopted legislation*'), the Singapore government has ordered individuals and organisations to post correction notices next to content that is deemed false (Palma, Munshi & Reed, 2020).

The law allows the possibility of appeal in court. In January 2020, the opposition party, the Singapore Democratic Party, filed the first appeal (Singapore Democratic Party, 2020), but this was dismissed by the High Court (AFP, 2020).

89. Sri Lanka

In April 2019, a terrorist attack on churches and hotels on Easter Sunday resulted in anti-muslim violence in Sri Lanka. Social media sites were blocked in the days following the attack with a view to limiting incitement to violence against muslims (Ellis-Petersen, 2019). The *Washington Post* reported that similar measures were taken in Sri Lanka after anti-muslim violence erupted in 2018 (Romm, Dwoskin & Timberg, 2019).

90. Thailand

The Thai government has widely enforced the Computer Crime Act (*see* '*Adopted legislation*') to arrest individuals for spreading 'fake news' online (AFP, 2019b; Bangkok Post, 2019).

91. Ukraine

In 2017 and 2018, Ukraine restricted the Yandex search engine, the social-media networks VKontakte and Odnoklassniki, and another 192 websites in Ukraine, as national security measures and economic sanctions against the Russian Federation (Oliphant, 2017; Jankowicz, 2019).

Photo credits: Shutterstock, igorstevanovic

#DISINFODEMIC
#THINKBEFORESHARING
#SPREADKNOWLEDGE

BROADBAND COMMISSION
FOR SUSTAINABLE DEVELOPMENT

9 789231 004032