Discussion Paper

on

An Australian Voluntary Code of Practice for Disinformation

Prepared for Digital Industry Group Inc. (DIGI) by UTS Centre for Media Transition

16 October 2020

Contents

Introduction: Overview of the project

This Discussion Paper provides background research relevant to the development of a voluntary code of practice for disinformation and is a companion document to the draft Australian Code of Practice on Disinformation ('the Code') being released for public consultation. Over time, digital platforms have introduced measures to counter disinformation and enable the public to make informed decisions in relation to content; the Code provides an opportunity to develop a common set of principles and commitments in relation to this work by platforms and to build on existing efforts.

The development of this Code has been driven by the Digital Industry Group Inc. (DIGI). DIGI is a non-profit industry association that advocates for the interests of the digital industry in Australia, with Google, Facebook, Twitter and Verizon Media as its founding members. DIGI also has an associate membership program and our other members include Redbubble, eBay, GoFundMe and Change.org. DIGI's vision is a thriving Australian digitally enabled economy that fosters innovation, a growing selection of digital products and services, and where online safety and privacy are protected.

DIGI commissioned the Centre for Media Transition (CMT) at University of Technology Sydney to assist with the preparation of the Code and the Discussion Paper. CMT, an interdisciplinary research centre that investigates key areas of media evolution and digital transition, drew on the assistance of First Draft, a global organisation that empowers societies with the knowledge, understanding and tools needed to outsmart false and misleading information.

This work is being undertaken as part of DIGI's response to Government policy as set out in *Regulating in the Digital Age: Government Response and Implementation Roadmap for the Digital Platforms Inquiry*, developed following the ACCC's Digital Platforms Inquiry. The Roadmap states:

> The Government will ask the major digital platforms to develop a voluntary code (or codes) of conduct for disinformation and news quality. The Australian Communications and Media Authority (ACMA) will have oversight of the codes and report to Government on the adequacy of platforms' measures and the broader impacts of disinformation.
>
> The codes will address concerns regarding disinformation and credibility signalling for news content and outline what the platforms will do to tackle disinformation on their services and support the ability of Australians to discern the quality of news and information. The codes will be informed by learnings of international examples, such as the European Union Code of Practice on Disinformation. The Government will assess the success of the codes and consider the need for any further reform in 2021.

The project involves research on existing approaches to managing disinformation and consultation with the digital industry on platforms' own initiatives for addressing the problem.

This paper provides background and context to help industry participants, government and the community consider how they can work together to tackle the issue of disinformation and misinformation, while at the same time promoting the value of free speech in an open democratic society. It covers:

- the concept of disinformation and how it relates to misinformation
- relevant industry initiatives

- international initiatives – regulation in other jurisdictions.

By considering different ways of defining disinformation and various international approaches to regulation, we hope this paper helps to reveal the different dimensions of disinformation and some of the challenges for regulation.

1 Information disorder: the background for an Australian code of practice

Aim of this paper

There are complex issues that arise in approaching the topic of disinformation and misinformation. Foundational questions – such as what to regulate and who should be the subject of regulation – are being confronted internationally. Naturally, there are differing views about some of these matters, but in order to reach an effective and proportionate regulatory outcome, these views need to be considered. Some of the specific challenges involved in designing regulation in this area include:

- the choices that must be made in defining disinformation – including the type of 'harms' which are included within that concept;
- the risks to freedom of speech, including political communication, that may arise in the course of taking action in relation to content;
- the difficulties of setting regulatory initiatives at a national level for issues that affect a range of industry participants and consumers across multiple jurisdictions;
- the need to combine regulatory approaches with other initiatives to raise awareness and media literacy or to encourage factual accuracy in news reporting;
- how regulation can encourage a sense of shared responsibility among the community, government, content producers and digital platforms.

The aim of this paper is to inform discussion about the complexities and potential challenges when responding to online disinformation in the Australian regulatory context. It seeks to explore these issues and provide some background and guidance for DIGI in developing its voluntary industry code of practice.

A threshold challenge is identifying what constitutes disinformation and differentiating it from other content and conduct which is the subject of regulation.

Approaches to the regulation of speech, including online speech, vary across jurisdictions. In Australia, liability can arise in relation to content and conduct such as violent live-streamed material, cyberbullying and image-based abuse, defamation, misleading and deceptive content and even inaccurate news. The sources of regulation include national security legislation, criminal law, communications regulation and various forms of voluntary, industry-based regulation. These various forms of existing regulation show that responsibility for addressing content and conduct that can be harmful is likely to be shared across a range of participants. This includes, at various points, suppliers of communications access and infrastructure, services providers including digital platforms, content producers and users of online service. In the case of defamation law, for example, liability might be shared across a content creator such as a news media outlet, a social media service used to distribute the content, and users who post comments against news articles. As the Australian Government's *Implementation Roadmap* showed, regulation in this field is still evolving.

In the context of this multiplicity in sources of regulation, this section tries to identify disinformation as a distinct category of content or conduct. We preface this with some

additional context by briefly considering some research on community understanding of the problem in Australia. After giving an explanation about some of the complexities involved in deciding on appropriate terminology, we then look at some specific challenges in fashioning a definition of disinformation, before outlining some specific ways in which disinformation manifests, including in Australia. We end this section by looking at some aspects of news credibility.

Australians' perceptions of information disorder

In the *Digital News Report* (DNR) for 2020, 64% of Australians reported high levels of concern about misinformation.[1] This has been consistently high since 2018, and ranks Australia as the tenth most concerned nation out of 40 markets surveyed.[2] This concern is predominantly around misinformation on social media, where strong perceptions of 'the prevalence of fake news'[3] correlate to notably low levels of trust in news sourced on digital platforms.[4] Despite steady increases in the use of social media to access news, nearly half who consume news this way do not trust what they see.

From multiple surveys a broad picture emerges of who is most concerned about misinformation and how they perceive it. Despite lower levels of engagement with news on social media, older generations express the highest levels of concern.[5] Despite having less concern about misinformation, younger generations are more likely to fact-check news accessed online.[6] This speaks more to higher levels of literacy for 'digital native' generations.

People who already pay for their news, as well as people who have a strong interest in politics, are two other groups more likely to be concerned about misinformation.[7] When asked about what kinds of misinformation concerns them most, Australians feel that political misinformation produced by the government, politicians or political parties is the highest concern.[8] However, the DNR survey data suggests political orientation plays a considerable role in how misinformation is perceived. People who identify as left-wing, for example, are far more likely to be concerned about government and political misinformation,[9] whereas people

[1] Park, S. et al. 2020. *Digital News Report: Australia 2020*. Canberra: News and Media Research Centre. See https://apo.org.au/node/305057 p 77.

[2] Newman, M. et al. 2020. *Reuters Institute Digital News Report 2020*. Oxford: Reuters Institute for the Study of Journalism. See https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf, p 18

[3] Ipsos Global Advisor, 2019. *Ipsos Trust in the Media*. See https://www.ipsos.com/sites/default/files/ct/news/documents/2019-06/global-advisor-trust-in-media-report-24jun2019.pdf

[4] Flew, T. & Dulleck, U. & Park, S. & Fisher, C. & Isler, O. 2020. 'Trust and Mistrust in News Media' *Best Centre Report, Queensland University of Technology*. See https://research.qut.edu.au/best/wp-content/uploads/sites/244/2020/03/Trust-and-Mistrust-in-News-Media.pdf p12.

[5] Park, S. et al. 2020. *Digital News Report: Australia 2020*. Canberra: News and Media Research Centre. p12 https://apo.org.au/node/305057 p 78.

[6] Fisher, C. et al. 2019. *Digital News Report: Australia 2019*. Canberra: News and Media Research Centre. See https://apo.org.au/node/240786, p 90.

[7] Ibid p 86.

[8] Park, S. et al. 2020. *Digital News Report: Australia 2020*. Canberra: News and Media Research Centre. See https://apo.org.au/node/305057 p 79.

[9] Ibid p 80

who identify as right-wing are most concerned about activist groups and activists spreading misinformation.[10]

In the experience of many Australians online, concern does not always translate to action. Australians tend to take a more passive approach to misinformation and usually will not verify information they are accessing through social media. While some people will feel hesitant about sharing news they are suspicious of, this hesitancy does not mean suspicious information is discounted altogether, and some may still consider sharing it.[11] Research shows that this is particularly the case among those with a lower interest in news; lower education levels; and among older generations.[12]

When asked what should be done about misinformation, Australians lean towards several solutions. The Australia Institute noted there is a level of responsibility ascribed to political parties with '84% of Australians supporting truth in political advertising laws – a result which held across all political persuasion'.[13] The 2020 *Digital News Report* noted 58 per cent of those surveyed think 'it is up to the tech companies to "block" those responsible for the posts'.[14] The issue of 'fake news' merited a different response, with one 2018 study showing Australians very strongly felt that misinformation around poor journalism lay with media companies and journalists ahead of digital platforms.[15]

These findings indicate the importance of media literacy efforts for an Australian code of practice. Given the challenges explored in this section in relation to dis- and misinformation, and the potential silencing of freedom of expression and political speech, strategic partnerships and initiatives in conjunction with the digital industry may be the most scalable solutions in such a complex area. In Section 3, we explore some of the initiatives in media literacy that are already being carried out by industry.

The problem of terminology

The problem of false information existed well before the digital era. Claire Wardle and Hossein Derakhshan put this into current day context in the 2017 Council of Europe Report (COE) *Information Disorder: Towards an Interdisciplinary Framework*:

> Politicians have forever made unrealistic promises during election campaigns. Corporations have always nudged people away from thinking about issues in particular ways. And the media has long disseminated misleading stories for their shock value. However, the complexity and scale of information pollution in our digitally connected world presents an unprecedented challenge.[16]

[10] Ibid p 80

[11] Ibid p 89

[12] Ibid pp 90-91

[13] See https://www.tai.org.au/content/truth-political-advertising-its-time-has-come.

[14] Above n 8

[15] Ibid p 38

[16] See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Finding salient and meaningful terms to define the issues remains complex,[17] and 'difficult'.[18] First Draft, the international news verification organisation, uses the term 'information disorder' to cover the types, phases, and elements of mis- and disinformation within the wider framework of the digital ecosystem:

> While the historical impact of rumours and fabricated content have been well documented, we argue that contemporary social technology means that we are witnessing something new: information pollution at a global scale; a complex web of motivations for creating, disseminating and consuming these 'polluted' messages; a myriad of content types and techniques for amplifying content; innumerable platforms hosting and reproducing this content; and breakneck speeds of communication between trusted peers.[19]

To help bring order to this environment, First Draft distinguishes between disinformation, misinformation and malinformation.[20]

Disinformation is false information that is deliberately created or disseminated with the express purpose to cause harm. Producers of disinformation typically have political, financial, psychological, or social motivations.

Misinformation is information that is false, but not intended to cause harm. For example, individuals who don't know a piece of information is false may spread it on social media in an attempt to be helpful.

Malinformation is genuine information that is shared to cause harm. This includes private or revealing information that is spread to harm a person or reputation.

These definitions have been adopted by UNESCO.[21] Wardle and Derakhshan's framework for managing the complex and multifaceted issues and overlaps of information disorder have been detailed in the Appendix of this paper. This framework outlines the types, phases and elements of information disorder, and takes into account the cycles of creation of content and re-creation and re-distribution of the content. But variations have been made, and efforts to combat information disorder have focussed on different elements. As noted above, in its 2019 policy announcement the Australian Government nominated 'disinformation' along with 'credibility signalling for news content' as the aspects that should be addressed by industry in the voluntary code. Then, when releasing a position paper on the issue in June 2020, the ACMA used the umbrella term 'misinformation' to describe these various manifestations of information disorder.[22] In considering how to approach the various elements that must be considered in connection with disinformation, we first look at the example of the EU Code of Practice on Disinformation.

---

[17] Claire Wardle. 'Fake news. It's complicated.' See https://firstdraftnews.org/latest/fake-news-complicated/.

[18] 'Tackling misinformation in an open society: How to respond to misinformation and disinformation when the cure risks being worse than the disease'. Full Fact 2018. See https://fullfact.org/blog/2018/oct/tackling-misinformation-open-society/.

[19] See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c p 4.

[20] These definitions are taken from First Draft's *Information Disorder: The Essential Glossary*, July 2018. See https://firstdraftnews.org/wp-content/uploads/2018/07/infoDisorder_glossary.pdf?x46415 .

[21] See https://en.unesco.org/sites/default/files/journalism_fake_news_disinformation_print_friendly_0.pdf.

[22] While it references the work of First Draft, the ACMA uses 'misinformation' as its collective term. See *Misinformation and news quality on digital platforms in Australia: A position paper to guide code development*, June 2020, p11-12.

The EU Code of Practice as a reference point for a self-regulatory code

The EU Code of Practice on Disinformation is the principal self-regulatory instrument that has been developed to tackle disinformation on digital platforms. Most of the potential signatories to the Australian Code are businesses that operate internationally – and some of them have already made commitments in keeping with the EU Code. For these reasons, it is an important reference point in the development of an Australian code of practice.

The EU Code defines disinformation as follows.[23]

EU Code of Practice on Disinformation
The Code adopts the definition used in the European Commission Communication 'Tackling online disinformation: a European approach'. The Code defines disinformation as, '"verifiably false or misleading information" which, cumulatively, "is created, presented and disseminated for economic gain or to intentionally deceive the public"; and "may cause public harm", intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security."'[24] The Code clarifies what is *not* disinformation. Particularly, disinformation 'does not include misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary, and is without prejudice to binding legal obligations, self-regulatory advertising codes, and standards regarding misleading advertising.'

While the definition of disinformation in the EU Code is a useful point of reference, several of the concepts it embodies require specific consideration in the Australian environment. The following sections discusses some of the difficulties involved in formulating definitions.

How does disinformation fit with existing regulations?

An important aspect of the EU Code is that it is designed to apply across the various states of the EU, where various national laws continue to operate. When designing a code of practice for a single jurisdiction such as Australia, it is important to consider the extent to which the EU Code is an appropriate model. At a national level, it is easier to see how the subject matter covered by such a new regulatory instrument sits alongside other forms of regulation. Some forms of content and conduct will inevitably overlap – disinformation is sometimes a feature of hate speech, for example – and it will be important to consider how

[23] See https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[24] EU Code referencing: 'European Commission Communication 'Tackling Online Disinformation: A European Approach' paragraph 2.1. In paragraph (b), 'intended as' refers to the original definition in the Communication, which put it this way: 'Public harm comprises threats to democratic political and policy-making processes as well as …' See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.

such overlaps should be handled.

Is intention an element of disinformation?

Like First Draft, UK fact checking organisation Full Fact describes misinformation as the 'inadvertent spread of false or misleading information' and disinformation as 'the deliberate use of false or misleading information to deceive audiences'.[25]

The question of whether disinformation requires an element of intention needs careful consideration. First Draft acknowledges complexities arise, for example in the coronavirus pandemic, when people share harmful yet false health information or conspiracy theories and genuinely believe them to be true. Full Fact also notes it is not always helpful to 'divide the issues by intent', but rather to leave it up to the public to 'judge where inaccuracies lie on the spectrum of misinformation and disinformation.'[26] First Draft's framework (see Appendix) further expands on the complexities of the definitions by outlining the types, phases and elements of information disorder where the re-production of a message may be different from that of the original creator of the message.

In the EU Code intention appears only in relation to an intention to deceive, and harm is regarded objectively (i.e., the material or conduct may, as a matter of fact, cause harm) rather than subjectively (the actor intended to cause harm). A variation on this approach is seen in the UK Government's *Online Harms White Paper*, where disinformation is described as 'information which is created or disseminated with the deliberate intent to mislead; this could be to cause harm, or for personal, political or financial gain'. This approach applies an element of intention both to the act of misleading and to the causing of harm.[27] The difficulties in establishing intention may make it more appropriate to make conduct, rather than content, the focus of regulatory attention.

What kinds of harm are within scope of disinformation?

In addition, the concept of 'harm' is something that will need to be considered in relation to local standards and community expectations, and in recognition that the level of harm associated with disinformation varies greatly. The EU Code relies on the concept of 'public harm' which it defines as 'threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security'.

The need for careful consideration of the concept of harm is noted by Full Fact, which proposes five levels of harm ranging from 'risk to life', 'economic harm', 'interference in democracy', 'disengagement from democracy' and, finally, 'no harm'.[28] This framework helps us to understand the variation in levels of harm, and it is also useful is showing there will be limits to what can be achieved through regulation: while it is easy to see a role for regulation in helping to prevent risk to life, for example, it is less likely that regulation will have a direct

[25] 'Tackling misinformation in an open society: How to respond to misinformation and disinformation when the cure risks being worse than the disease'. Full Fact 2018. See https://fullfact.org/blog/2018/oct/tackling-misinformation-open-society/.

[26] Ibid

[27] Department for Digital, Culture, Media and Sport and Home Office, *Online Harms White Paper (*April, 2019). The paper distinguishes disinformation from misinformation which it describes as 'the inadvertent sharing of false information' (p 23). See https://assets.publishing.service.gov.uk/Online_Harms_White_Paper.pdf.

[28] Tackling misinformation in an open society: How to respond to misinformation and disinformation when the cure risks being worse than the disease. See. https://fullfact.org/blog/2018/oct/tackling-misinformation-open-society/ pp 5-7.

role in preventing disengagement from democracy.

The impact on political expression – protecting freedom of speech

In Australia, as in other liberal democracies, one of the most important contextual aspects for developing any rules or laws in relation to online content is the need to avoid imposing unnecessary restraints on freedom of speech and expression. While the community accepts that some forms of speech must be restricted – for example, child abuse material or image-based abuse – users of digital platforms also expect a degree of freedom in their ability to post their own content and access that of other users and content creators. Additional complexities arise in assessing and classifying political expression as disinformation,[29] for example when facts are hyperbolic or exaggerated in political expression or in media communications. At one end of the spectrum fact checkers may rightly challenge and correct figures about issues such as tax cuts;[30] and in other cases, fringe politicians in Australia have pushed extreme right wing anti-immigration sentiment.[31] But people may also simply disagree with opposing political statements and attempt (incorrectly) to label this as misinformation. It is important also to recognise that information is never 'perfect' and that factual assertions are sometimes difficult to verify. As Deborah Stone has noted, in democratic decision-making 'information is interpretive, incomplete, and strategically withheld' by participants, including political parties and government actors.[32] These challenges need careful consideration in a code of practice, particularly how commitments made under the code may be misused with the intention of silencing political opposition. The EU Code addresses this through a provision that companies should not be compelled by governments to remove content because of perceived falsity:

> Signatories should not be compelled by governments, nor should they adopt voluntary policies, to delete or prevent access to otherwise lawful content or messages solely on the basis that they are thought to be 'false'.

Satire and other forms of creative expression

Consideration needs to be given to whether creative expression should be excluded from a definition of disinformation. In satire, for example, a falsity or exaggeration might be used for humour to make a broader critique or a form of political or social expression. Satire is often a central feature of political cartoons in countries like Australia, so that cartoons are generally accorded a greater leniency under media standards that relate to offence, for example. Satire is expressly excluded from the EU Code. However, First Draft clarifies, that if a person takes a satirical news story literally for example, and shares it with this mistaken belief, that could be considered misinformation. Additionally, if satirical memes are used as part of a campaign to discredit a person or racial community, this could fall under defamation, hate speech, cyber-bullying or disinformation.

[29] See https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/.

[30] See https://blogs.lse.ac.uk/medialse/2019/12/12/online-political-advertising-in-the-uk-2019-general-election-campaign/.

[31] See https://firstdraftnews.org/latest/tracking-anti-muslim-tactics-online-australias-election-misinformation/.

[32] Deborah Stone (2002), *Policy Paradox: The Art of Political Decision-Making*, WW Norton, p 28.

Example

In February 2020, an Australian couple posted on Facebook they had ordered wine using a drone while they were quarantined on the Diamond Princess ship off the coast of Tokyo.[33] It was reported on by international media from Hong Kong[34] to the *New York Post* and was re-shared on Facebook and Twitter including by celebrities.[35] The couple later confirmed it was a joke for their friends, and commented that no reporter had checked with them until it was finally 'fact checked' by ABC Radio National.[36]

How information disorder manifests

Having considered some of the elements of disinformation and some aspects that will need to be taken into account for a definition of disinformation in the Australian Code, we turn now to look at how disinformation is propagated and the forms it takes. We consider some important elements, such as the role of malicious actors, and provide some specific examples relevant to Australia. We then look at the 'ABC' conceptual framework developed by Camille Francois, which explains the role of malicious actors, deceptive behaviour and harmful content.

During coronavirus, many seemingly disparate groups have used the heightened sense of awareness and fear from the public to promote conspiracy theories, vaccine hesitancy and encourage people who are clearly and unequivocally against vaccinations. First Draft has identified a significant increase in online activity among groups sharing anti-vaccination content in Australia since the start of COVID-19. The motivations of agents of disinformation, as well as the tools and techniques used in information disorder are outlined below.

Agents of disinformation are motivated broadly by power, money, or mischief. Wardle and Derakhshan[37] further specify the motivations as:

- Financial: profiting from information disorder through advertising;
- Political: discrediting a political candidate in an election and other attempts to influence public opinion;
- Social: connecting with a certain group online or off (this has ramifications for information disorder where people are 'recruited to an ideology' or join conspiracy

[33] AFP Fact Check, 'Australian couple quarantined onboard Diamond Princess cruise reveal wine drone delivery story was "just a prank"' (February, 2020). *AFP* . See https://factcheck.afp.com/australian-couple-quarantined-onboard-diamond-princess-cruise-reveal-wine-drone-delivery-story-was.

[34] 9GAG,see https://perma.cc/RFC6-HT8D.

[35] https://twitter.com/Kate_Chastain https://perma.cc/CE8P-Z594.

[36] Paul Barry, 'Media Tricked' (February, 2020) *ABC* . See https://www.abc.net.au/mediawatch/episodes/drone/.

[37] See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c . Research methodology (page 7) noted: 'In addition to our conceptual framework, we provide a round-up of related research, reports and practical initiatives connected to the topic of information disorder, as well as filter bubbles and echo chambers. We examine solutions that have been rolled out by the social networks and consider ideas for strengthening existing media, news literacy projects and regulation. We also introduce some key future trends, particularly in terms of the rise of closed messaging apps and the implications of artificial intelligence technology for manufacturing as well as detecting dis-information.'

theory groups); and,

- Psychological: seeking prestige or reinforcement.

First Draft has adapted the work of Data & Society[38] to identify the methods and tools used by agents of disinformation. These have become more sophisticated and include:[39]

- Sockpuppet Accounts: the anonymous figures – bot, human, or hybrid – pretending to be something they are not.
- Imposter Content: using trusted logos, branding or names as a shortcut for credibility.
- Source Hacking: manipulating the news media and influential figures through lies and deception.
- Keyword Squatting: associating a word with a worldview.
- Information bombardment to overwhelm and confuse.

Example: Disinformation & public figures

In April 2020, the Therapeutic Goods Administration[40] launched an investigation into controversial celebrity chef Pete Evans after his Facebook Live video over Easter public holidays touted his 15,000 USD 'bio light'. He also used divisive language referring to the coronavirus as 'Wuhan coronavirus'. The celebrity chef's podcasts[41] alluded to 5G conspiracies and use of vitamins to ward off the coronavirus.

Example: Agents of disinformation take advantage

The hashtag #ArsonEmergency was first used in November 2019 at the same time #ClimateEmergency began trending during the first round of Australia's devastating summer of bushfires. #ArsonEmergency did not pick up in usage until early 2020 when the researchers found it was pushed in a sustained effort by around 300 inauthentic accounts.[42] From here, it was adopted by genuine accounts as the narrative was pushed further into mainstream conversation. As AFP fact-check pointed out, the arson claim was published widely across conservative news outlets including The Australian;[43] The Sun (UK); and Breitbart (US).[44]

[38] See https://datasociety.net/library/source-hacking-media-manipulation-in-practice/.

[39] See https://firstdraftnews.org/en/education/curriculum-resources/.

[40] See https://www.tga.gov.au/.

[41] See https://www.youtube.com/watch?v=oZtWCMVk_GY.

[42] Esther Chan, 'Debunking the bushfires disinformation inferno' (February, 2020) AFP. See <https://correspondent.afp.com/debunking-bushfires-disinformation-inferno>.

[43] See https://www.theaustralian.com.au/nation/bushfires-firebugs-fuelling-crisis-asarson-arresttollhits183/news-story/52536dc9ca9bb87b7c76d36ed1acf53f.

[44] AFP Fact Check, 'Police figures show far fewer people in Australia have been charged with bushfire arson' (January, 2020) *AFP.* See https://factcheck.afp.com/police-figures-show-far-fewer-people-australia-have-been-charged-bushfire-arson.

(cont.)

Confusion over the term 'arson' was further exacerbated in early January after The Australian reported more than 180 alleged arsonists had been arrested since the start of 2019.[45] This, and many other headlines misconstrued a New South Wales Police force media release. As Vox quickly reported in order to debunk the story: '[w]hat the release actually says is that legal action was taken against 183 people since November 8, 2019, for fire-related offenses, including things like improperly discarding cigarettes or not taking enough precautions around machinery, i.e. not arson.'[46] The false claim was picked up and amplified on the international stage by Donald Trump Jr., Fox News, famous alt right figures and websites. A Google search for 'Australia and bushfires in that same week returned headlines focused on the 'arson crisis' topic and pitched this to question climate change. However, as debunks filled the 'data voids', more reliable stories quickly showed up higher in the search results.

The term 'data voids', created by danah boyd and Michael Golebiewski, provides a useful concept for journalists reporting on disinformation.[47] Examples from the case studies point to the importance of journalists and platforms working together to address public questions and fill the voids with reliable information. For example, when Google searches show a spike for a particular term that has not surfaced before, or returns few meaningful results, this 'data void' provides an opportunity for content from bad actors to surface. So while a search of 'bushfire Australia' initially turned up 'Arson emergency' related stories, as ranking adjusted to more available quality information and as journalists corrected and replaced the topic with debunks, the search returned fact checked information first.

The role of malicious actors

Malicious actors act to inflict harm on a person, organisation or country. As methods can include leaks, harassment and hate speech, there is a clear overlap with other forms of

[45] See https://www.theaustralian.com.au/nation/bushfires-firebugs-fuelling-crisis-asarson-arresttollhits183/news-story/52536dc9ca9bb87b7c76d36ed1acf53f.

regulated speech. This conceptual distinction is useful as it helps to show how aspects of information disorder may already be the subject of existing regulation, including, in some cases, criminal law. This is particularly the case in relation to malicious actors and the spread of malinformation.

In explaining malinformation, Claire Wardle noted, 'It is important to distinguish messages that are true from those that are false, but also those that are true (and those messages with some truth) but which are created, produced or distributed by 'agents' who intend to harm rather than serve the public interest.'[48] For example, The Mueller Report established the social media campaign by Russian actors included a hacking operation against the Clinton Campaign which released stolen documents.[49]

Australia has not seen high profile public examples of malinformation when compared to examples arising out of the US; however, the subject of foreign interference in elections and in democratic formations more generally has been the subject of political inquiry.[50] The risk of this occurring in future in Australia could overlap with national security concerns with campaigns by foreign agents, however little details are available publicly on this issue. In February 2019, three months ahead of the federal election, Canberra confirmed government computers had been hacked, and described the level of sophistication as 'unprecedented'.[51] [52] The government did not disclose which country they believed was responsible. In September 2019, Reuters reported that anonymous sources from the Australian Signals Directorate (ASD) found in March that China's Ministry of State Security was responsible for the hack on MPs' emails.[53] This included the networks of the Australian Labor Party, the Liberals and the Nationals.[54] While this behaviour may be regarded as

[46] Umair Irfan, 'The viral false claim that nearly 200 arsonists are behind the Australia fires, explained' (January, 2020) *Vox.* See https://www.vox.com/2020/1/9/21058332/australia-fires-arson-lightning-explained.

[47] See https://datasociety.net/library/data-voids/.

[48] Clare Wardle and Hossein Derakhshan, 'Journalism, "Fake News" & Disinformation' (2018) *UNESCO* <https://en.unesco.org/sites/default/files/f._jfnd_handbook_module_2.pdf> p 44.

[49] Robert S Mueller III, 'Report On The Investigation Into Russian Interference In The 2016 Presidential Election' (March, 2019) Volume I. See https://www.justice.gov/storage/report.pdf p 4.

[50] For example, in 2018 the Parliamentary Joint Committee on Intelligence and Security (PJCIS) completed a Review of the National Security Legislation Amendment (Espionage and Foreign Interference) Bill 2017. The PJCIS has also been asked by the Minister for Home Affairs to conduct an inquiry into foreign interference in Australia's universities, publicly funded research agencies and competitive research grants agencies. See https://www.aph.gov.au/About_Parliament/House_of_Representatives/About_the_House_News/Media_Releases/Foreign_interference_in_universities_inquiry_under_consideration. In addition, the Senate Select Committee on Foreign Interference through Social Media is to report by May 2022. See https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Foreign_Interference_through_Social_Media/ForeignInterference.

[51] David Wroe and Chris Uhlmann, 'Australia's major political parties hacked in "sophisticated" attack ahead of election' (February, 2019) *Sydney Morning Herald*. See https://www.smh.com.au/politics/federal/australia-s-major-political-parties-hacked-in-sophisticated-attack-ahead-of-election-20190218-p50yi1.html.

[52] See https://www.reuters.com/article/us-australia-china-cyber-exclusive/exclusive-australia-concluded-china-was-behind-hack-on-parliament-political-parties-sources-idUSKBN1W00VF.

[53] Colin Packham, 'Exclusive: Australia concluded China was behind hack on parliament, political parties – sources' (September, 2019) *Reuters.* See https://www.reuters.com/article/us-australia-china-cyber-exclusive/exclusive-australia-concluded-china-was-behind-hack-on-parliament-political-parties-sources-idUSKBN1W00VF.

[54] Rob Harris, 'Intelligence agencies pinned Parliament hack on Beijing: report' (September, 2019) *Sydney Morning Herald.* See https://www.smh.com.au/politics/federal/intelligence-agencies-pinned-parliament-hack-on-beijing-report-20190916-p52rou.html.

espionage or some other form of offence against national security, hacked emails that are leaked and framed negatively can also be regarded as a form of malinformation.

A disinformation 'ABC'

Camille Francois from Graphika and the Berkman Klein Center for Internet & Society at Harvard University has distilled definitions of disinformation into what is known widely in the research community as an 'ABC' framework focused on actors, behaviour and content.[55] Francois noted that 'manipulative actors' (with the clear intention to disrupt the information ecosystem), 'deceptive behaviors' (tactics and techniques used by the actors) and 'harmful content' (used to hurt, undermine or influence) are 'three key vectors characteristic of viral deception'.[56]

'A': manipulative actors

Manipulative actors 'engage knowingly and with clear intent in viral deception campaigns.'[57] The actors' intent and their campaigns are 'covert, designed to obfuscate the identity and intent of the actor orchestrating them'.[58] Russian disinformation campaigns that targeted the US 2016 presidential election provide an example of covert actors with the intent to deceive. The Mueller Report established the intent of the social media campaign by Russian actors 'favored presidential candidate Donald J. Trump and disparaged presidential candidate Hillary Clinton.'[59] Investigations for the Mueller Report showed that the Internet Research Agency (IRA), based in St Petersburg, Russia, 'carried out the earliest Russian interference operations', and 'received funding from Russian oligarch Yevgeniy Prigozhin' who 'is widely reported to have ties to Russian President Vladimir Putin.'[60] The campaign involved a social media campaign designed to 'provoke and amplify political and social discord in the United States' and a hacking operation against the Clinton Campaign which released stolen documents.[61]

For an example of manipulative actors in the Asia region, harassment and hate speech erupted on social media platforms commenting on the 2019 Hong Kong summer protests. This fuelled polarisation and limited constructive dialogue. On August 20, 2019, Twitter identified 'that Chinese State-controlled media leveraged Twitter advertising to promote content critical of pro-democracy protests in Hong Kong.'[62]

Cybersecurity concerns in Australia is a potential overlapping area which risks exposure to

---

[55] *Actors, Behaviors, Content: A Disinformation ABC. Highlighting Three Vectors of Viral Deception to Guide Industry and Responses*. A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of expression. Graphika and Berkman Klein Center for Internet & Society at Harvard University p 2.

[56] Camille Francois, 'Actors, Behaviours, Content: A disinformation ABC'; See also: Kathleen Hall Jamieson, *Cyberwar: How Russian Hackers and trolls Helped Elect a President What We Don't, Can't, and Do Know. (*Oxford University Press, 2018).

[57] Ibid p 2.

[58] Ibid.

[59] Robert S Mueller III, *Report on the Investigation into Russian Interference in the 2016 Presidential Election* (March, 2019) Volume I. See https://www.justice.gov/storage/report.pdf p 1.

[60] Ibid p 4.

[61] Ibid.

[62] Above n 56.

disinformation campaigns led by manipulative actors. The Australian Strategic Policy Institute identified foreign interference in 20 countries including Australia.[63] On June 19, 2020, the Prime Minister Scott Morrison delivered a public address where he announced 'Australian organisations are currently being targeted by a sophisticated state-based cyber-actor.'[64] The risks were earlier acknowledged when, on December 5, 2019, the Senate established the Select Committee on Foreign Interference through Social Media to inquire into and report on the risk posed to Australia's democracy by foreign interference through social media.

Detection of state-based manipulative actors has traditionally been the domain of cyber-security government departments. Francois noted, however, that 'disinformation actors exploit the whole information ecosystem'[65] and called for a more integrated response to reach across all products and services offered by platforms. Given that this exploitation will include various tools, techniques and content, it is likely this may overlap with mis- and disinformation.

This means that security measures that various platforms have adopted serve to ensure the integrity of their services and authenticity of their users' accounts to prevent manipulative actors are therefore relevant actions to address disinformation, and should be considered under an Australian code of practice. Current platform interventions in relation to preventing fake accounts, and measures aimed at preventing impersonation, are explored in Section 2.

'B': deceptive behaviour

The next 'vector of disinformation' in Francois' 'ABC' framework is known as 'Deceptive Behavior' which is focused on the techniques used by deceptive actors.[66] The goal of these techniques is to give the impression of a greater impact as if there were larger numbers of actors. These techniques range from 'automated tools (e.g., bot armies used to amplify the reach and effect of a message) to manual trickery (e.g., paid engagement, troll farms).'[67]

Francois noted, 'while there are significant differences in the various disinformation definitions and terms of service applicable to the issue among technology companies, the focus on deceptive behavior appears to be a clear convergence point throughout the technology industry.'[68] For example, Google's February 2019 White Paper, *How Google Fights Disinformation*, noted:

> ... the words 'misinformation', 'disinformation', and 'fake news' mean different things to different people and can become politically charged when they are used to characterize the propagators of a specific ideology or to undermine political adversaries.

[63] Australian Strategic Policy Institute, Policy Brief, Report No. 16/2019 'Hacking democracies: cataloguing cyber-enabled attacks on elections'. See https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2019-05/Hacking%20democracies_0.pdf?.RKLLc8uKm1wobfWH1VvC.C88xGWYY29>.

[64] Daniel Hurst, 'Cyber-Attack Australia: sophisticated attacks from 'state-based actor', PM says' (June, 2020) *The Guardian*. See https://www.theguardian.com/australia-news/2020/jun/19/australia-cyber-attack-attacks-hack-state-based-actor-says-australian-prime-minister-scott-morrison.

[65] Above n 56 p 2.

[66] Ibid p 4.

[67] Ibid.

[68] Ibid.

> However, there is something objectively problematic and harmful to our users when malicious actors attempt to deceive them. It is one thing to be wrong about an issue. It is another to purposefully disseminate information one knows to be inaccurate with the hope that others believe it is true or to create discord in society.[69]
>
> We refer to these deliberate efforts to deceive and mislead using the speed, scale, and technologies of the open web as 'disinformation'.

The entities that engage in disinformation have a diverse set of goals. Some are financially motivated, engaging in disinformation activities for the purpose of turning a profit. Others are politically motivated, engaging in disinformation to foster specific viewpoints among a population, to exert influence over political processes, or for the sole purpose of polarizing and fracturing societies. Others engage in disinformation for their own entertainment, which often involves bullying, and they are commonly referred to as 'trolls'.[70]

Francois noted Facebook mostly defines deceptive behaviour through its 'Coordinated Inauthentic Behavior' policy.[71] Nathaniel Gleicher, Head of Cybersecurity Policy explained in December 2018:

> Coordinated Inauthentic Behavior is when groups of pages or people work together to mislead others about who they are or what they're doing. Coordinated Inauthentic Behavior isn't unique to Facebook, or social media. People have been working together to mislead others for centuries, and they continue to do so.
>
> When we take down one of these networks, it's because of their deceptive behavior, it's not because of the content they're sharing. The posts themselves may not be false and may not go against our community standards. We might take a network down for making it look like it's been run from one part of the world, when in fact it's been run from another.
>
> This could be done for ideological purposes, or it could be financially motivated, for example spammers might seek to convince people to click on a link to visit their page or to read their posts.[72]

Francois noted the 'detection and mitigation techniques' in deceptive behaviour can be 'similar to spam detection'.[73] As outlined above, platforms have proactive measures to identify problematic accounts and behaviours. Google utilises algorithmic signals to indicate deceptive behaviour. Where there is an indication that a publisher may be violating their policies, such as through a user report or suspicious account activity, Google's Trust and Safety team investigates and then, where appropriate, acts against that site and any related sites that can be confirmed to be operating in concert. Facebook utilises machine learning and AI in 'proactive' detection and take-down of Coordinated Inauthentic Behaviour - networks of accounts or pages working to mislead others about who they are, and what they are doing. Machine based learning is utilised by Twitter to identify and track accounts

---

[69] Google, 'How Google Fights Disinformation' (February, 2019). See https://kstatic.googleusercontent.com/files/388aa7d18189665e5f5579aef18e181c2d4283fb7b0d4691689dfd1bf92f7ac2ea6816e09c02eb98d5501b8e5705ead65af653cdf94071c47361821e362da55b 2.

[70] Ibid p 2.

[71] Above n 56 p 4.

[72] Nathaniel Gleicher, 'Coordinated Inauthentic Behavior explained' (December , 2018) *Facebook Newsroom*. See https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

[73] Above n 55 pp 4-5.

engaged in manipulative behaviour in order to identify inauthentic behaviour and neutralise this before users are exposed to misleading, inauthentic, or distracting content. LinkedIn utilises automated, machine learning to identify characteristics of bad actors and fake profiles.

Francois also noted that enforcement actions available to platforms such as content demotion and account suspension are 'rarely spelled out for users or made clear for users affected' and when it comes to manipulative actors and deceptive behaviour vectors, 'platforms have much more visibility into those issues than external researchers and stakeholders.'[74] However, platforms have noted the risk of their systems being gamed and exploited drawing upon any information that is made public about such actions, which can result in more deceptive behaviour. This risk must also be considered in Australian code of practice.

'C': harmful content

Francois noted content can lead to posts and messages being classified as viral deception, and is the most 'visible vector of the three: while it is difficult for an observer to attribute messages to a manipulative actor or to observe behavior patterns across a campaign, every user can see and form an opinion on the content of social media posts.'[75] Moderation of such content can intersect and overlap with other regulatory and legal frameworks e.g., 'harmful content', which Francois noted is the subject of ongoing debates about definitions including 'violent extremism', 'hate speech,' 'terrorist content'.[76] Francois noted 'entire categories of content can be deemed "harmful" because they belong to the realm of viral deception, eg, health misinformation'.[77] Additional ways Francois noted the intersection of harmful content and disinformation campaigns can manifest include:

- The content of a campaign itself can be manipulated to deceive users and therefore belong to the realm of 'disinformation', and,
- 'Harmful content' can be promoted by deceptive actors or by campaigns leveraging distortive behaviors.[78]

All platforms have policies in place to address the content, and the industry measures to address these issues are outlined in Section 3.

News credibility

In the final part of this section, we consider the role of traditional media. This is necessary in order to understand how disinformation might spread, but also how it might be addressed. In addition, as noted above, 'credibility signalling for news content' is an aspect that the Federal Government would like to see addressed in an Australian Code.

[74] Ibid p 5.

[75] Above n 55 p 6.

[76] Ibid p 6.

[77] Ibid p 6.

[78] Above n 55 p 6.

The 2018 report, *The Oxygen of Amplification*, noted, 'it is problematic enough when everyday citizens help spread false, malicious, or manipulative information across social media. It is infinitely more problematic when journalists, whose work can reach millions, do the same.'[79] This suggests that professional journalists and news organisations have a responsibility to be aware of the role they play in spreading and amplifying falsehoods.[80] First Draft uses the 'tipping point' to help journalists assess amplification risks where newsrooms must balance the public interest in the story against the possible consequences of overage. Misinformation or poor reporting by media also provides an opportunity for agents of disinformation to take advantage of the situation – see the example below, 'Agents of disinformation take advantage' with the case of the hashtag #ArsonEmergency in Australia. Media manipulation – where the goal of agents of disinformation is to have their issue reported on - is another consideration that journalists must be on guard for. Alice Marwick and Rebecca Lewis noted in their 2017 report, *Media Manipulation and Disinformation Online*, that, for manipulators, 'it doesn't matter if the media is reporting on a story in order to debunk or dismiss it; the important thing is getting it covered in the first place.'[81]

Fact checking

Accuracy is considered a core value shared by professional journalists, with the process of verifying facts a deliberate, conscious step for journalists, rather than it being an incidental by-product from gathering facts.[82] However it has been argued that 'many journalism textbooks are devoid of references to verification or fact-checking… or make only the briefest references to the importance of double-checking basic facts.'[83]

The digital era led to a global increase in the number of in-house fact checking units at media organisations in the US such as FactCheck.org which launched in 2003, and PolitiFact and the Washington Post's Fact Checker, which debuted in 2007. These fact checking units can have different foci – from checking 'the accuracy of the substantive claims made by politicians' rather than journalists simply trying to copy the quote from the politicians correctly[84] – to replicating the work of a media organization's legal department. Most prominently in Australia, ABC Fact Check was launched in 2013 to 'test and adjudicate on the accuracy of claims made by politicians, public figures, advocacy groups and institutions engaged in public debate',[85] and was re-launched as 'ABC RMIT FactCheck' in

[79] Whitney Phillips, Syracuse University, 2018 Data & Society report: *The Oxygen of Amplification: Better Practices for Reporting on Extremists*, Antagonists and Manipulators. See https://datasociety.net/library/oxygen-of-amplification/.

[80] First Draft Training: 'What does responsible reporting mean in an age of information disorder?'. See https://firstdraftnews.org/training/responsible-reporting/.

[81] Alice Marwick and Rebecca Lewis, 2017 *Media Manipulation and Disinformation Online*. Data & Society Research Institute. See https://datasociety.net/library/media-manipulation-and-disinfo-online.

[82] Kruger, A. L. (2019). *Ahead of the e-curve: Leading global social media verification education from Asia in a 21st century mediascape*. (Thesis). University of Hong Kong, Pokfulam, Hong Kong SAR. See http://hub.hku.hk/handle/10722/278427.

[83] Shapiro, I. Brin., Bedard-Brule, I., & Mychajlowycz,K. (2013). 'Verification as a Strategic Ritual'. Journalism Practice, *Journalism Practice*, 7 (6), 657-673, p 658.

[84] Stassen, 2010, 'Your News in 140 characters: Exploring the role of social media in journalism'. *Global Media Journal, African Edition*, 4 (1) p 118.

[85] See https://www.rmit.edu.au/news/all-news/2017/feb/rmit-and-abc-news-relaunch-fact-check.

2017. PolitiFact expanded to Australia in 2013[86] also with a focus on political fact checking.

However, media organisations in Australia (and globally) were left exposed to inaccuracies arising from UGC included or used as content in professional media reports. The BBC first adapted social media into its journalistic production practices in something of a 'trial and error' manner – potentially harming the reputation the BBC had with its audience.[87] In Australia, there are numerous cases where the media have not checked the provenance of online content and identities before publication.[88] This highlights the importance for journalists to understand the production, dissemination and interaction of messages and information in social media, so that news organisations can deliver reliable information for society. This required new skills to track and monitor social media.

In 2017 First Draft and Full Fact noted:

> fact-checking and verification have occupied quite different spaces within journalism, and the skills have been seen as distinct and specialist. Only with the rise of fabricated news websites did fact-checking and verification organizations find themselves both being asked how to 'debunk' these sites.[89]

The necessity of these skills, and knowledge of the tools and techniques to combat information disorder have since grown. While advanced verification training by organisations such as First Draft has focused on journalists, it has become increasingly evident that as celebrities and others who have a powerful platform have amplified and spread falsehoods, they too could well benefit society if they had some form of training in the principles of verification.

In March 2017, the then director of the International Fact Checking Network, Alexios Mantzarlis posted a tweet with the Venn diagram below, in an attempt to explain the relationship between fact-checking, verification and debunking.[90]

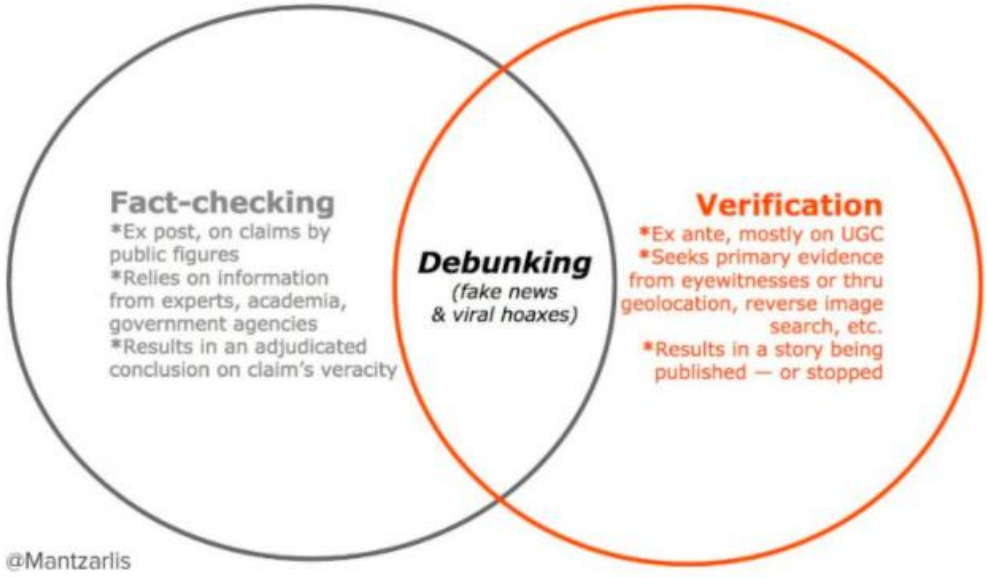
[86] See https://www.politifact.com/article/2013/may/12/politifact-expands-australia/.

[87] Belair-Gagnon, 2012, 'Getting it Right! : How did social media transform BBC News journalism' *Communiquer dans un monde de norms* p 237.

[88] Kaur, Kanchan and Nair, Shyam and Kwok, Yenni and Kajimoto, Masato and Chua, Yvonne T. and Labiste, Ma. Diosa and Soon, Carol and Jo, Hailey and Lin, Lihyun and Le, Trieu Thanh and Kruger, Anne. *Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam* (October 10, 2018). See SSRN: https://ssrn.com/abstract=3134581 or http://dx.doi.org/10.2139/ssrn.3134581.

[89] Claire Wardle, First Draft and Will Moy, Full Fact 'Is that actually true? Combining fact-checking and verification for #GE17. See https://firstdraftnews.org/latest/fullfact-ge17/.

[90] Ibid.



The role of traditional media

Misleading content frames information, an issue or an individual in a misleading manner. There are many instances of how this can occur. The following example shows how the role of news outlets can further amplify misinformation into online spaces, and lead to harmful assumptions which can open the way for agents of disinformation to use this to push their own agenda. Regulation that encourages verification training and media literacy awareness for the public would help to address the situation. Ongoing longitudinal research into the efficacy of the measures is also currently limited and would help to inform the design of such training.

Example: **Traditional media amplification**

In mid-February, major news organisations around the world published stories about a 'terrifying' map, with red lines crisscrossing and encircling the globe, lines they falsely claimed were how COVID-19 would spread, or how it had spread already.[91] Their source, they all reported, was a study from the WorldPop Project at the University of Southampton in the UK. The study, which was not peer-reviewed, estimated how many people had left Wuhan before the region was locked down. The image WorldPop initially tweeted to accompany the study however, showed global air-traffic routes and travel for the entirety of 2011. The tweet was hastily deleted with little explanation as to why, and the study reshared without the old image. But it had already been misinterpreted and republished by tabloids and television producers around the world including in Australia, apparently without any semblance of fact checking. The project responded to First Draft

[91] Carlotta Dotto and Jack Berkefeld, 'From coronavirus to bushfires, misleading maps are distorting reality' (February, 2020) *First Draft News.* See https://firstdraftnews.org/latest/from-coronavirus-to-bushfires-misleading-maps-are-distorting-reality/.

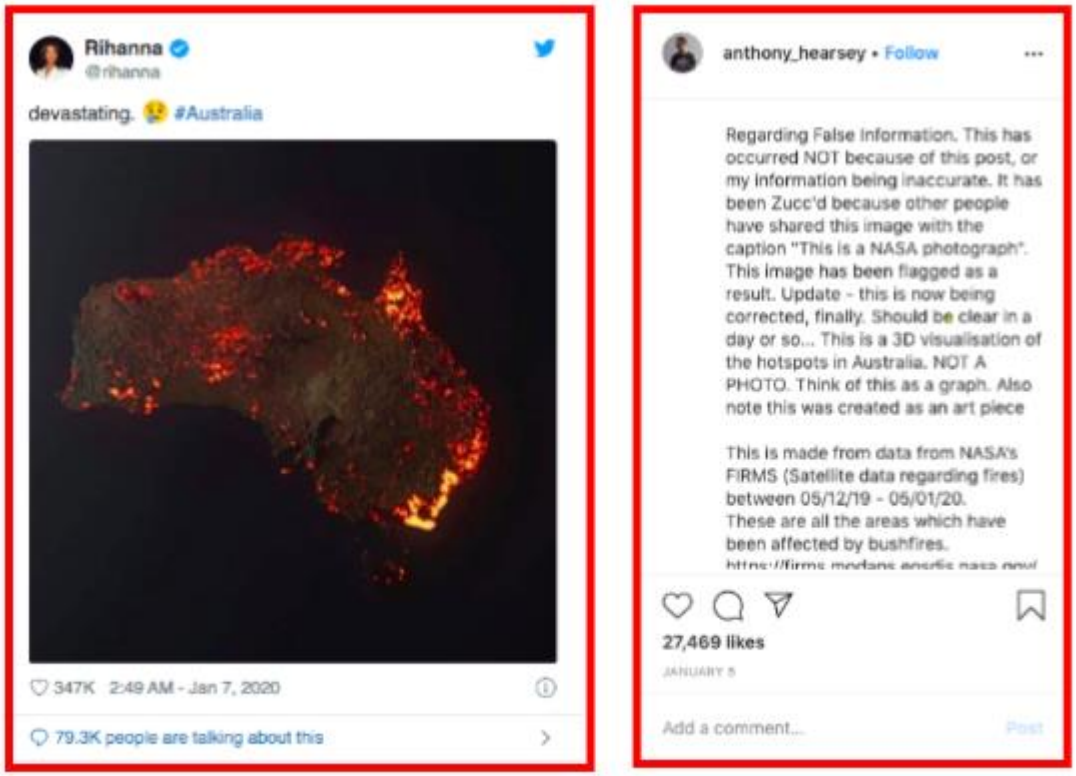
queries and described the image as 'intended to be an illustrative picture of the global air network'.[92] Such an image was easily misunderstood and required critical thinking. The image used in WorldPop's first tweet remains[93] (current as of July 13) on the official Channel 7 *Sunrise* @sunriseon7 twitter feed, and the video has been viewed 211, 600 times. This fuelled division and fear in a time of crisis.

Screenshot by Anne Kruger

While verification training has been aimed at journalists, recent examples show this should also be extended to public figures such as celebrities, politicians and sporting heroes; they too have a responsibility to be aware of the role they play in spreading and amplifying falsehoods due to their elevated reach.

[92] Ibid.

[93] See https://perma.cc/LS76-J45L Archive taken July 13,2020 of post from @sunriseon7 Twitter account.



The artist (r) pleaded with social media users not to share his work as accurate. Celebrity Rihanna (l) shared it with her 96 millions followers. Screenshots by author.

Example: Misinformation & public figures

At the height of Australia's extreme bushfire season, celebrity Rihanna tweeted a misleading picture of the country to her 96 million followers which enabled mass amplification. The 3D art was made using hotspot data from 31 days of fires according to the artist who created it, but it was mistaken for a NASA photograph. While the creator issued a public clarification on Instagram, this highlights questions of responsibility by original creators and re-sharers of online content. It also highlights as the possibilities for technological developments to aid audience understanding about the original source or purpose of the product.

2 Industry initiatives

Our research and consultations with industry have highlighted extensive efforts by platforms and providers of online content to encourage authenticity and transparency in online communication in relation to mis- and disinformation.

The common belief is that the challenge to address disinformation requires a holistic approach, with a recognition of responsibility and a range of measures and with varied responsibilities across the digital ecosystem. At the same time, providers have expressed a strong desire to not become 'the arbiters of truth', highlighting the complexities of this role in line with user expectations, and have arrived at technical and policy measures which both address the problems while also preserving the principles of freedom of expression.[94]

Additionally, platforms and online services continue to invest considerable resources in consumer programs to improve digital media literacy and fact-checking,[95] as well as efforts that inform research and counter emerging threats,[96] and journalism to support the news ecosystem.[97]

Industry consultations underline the scale of a task for operators in the digital space, and also the benefits of a unified approach in tackling mis- and disinformation. The unfolding global health crisis triggered by the COVID-19 pandemic, intensified the process of addressing disinformation and misinformation, and encouraged collaborative efforts by industry, researchers, government and public health agencies in seeking holistic measures to improve the quality and dissemination of COVID-19 information to online consumers.[98] Our consultations indicate platforms and providers recognise this period as an opportunity to progress their combined efforts in countering disinformation.

The diversity of digital products

The information ecosystem is comprised of a wide range of public digital resources demanding an equally diverse approach to resolving the impact of mis- and disinformation. In this complex network of search engines, software providers and user generated content platforms, each provider has its own unique set of functions, followers, and technical considerations to which no single technological fix, labelling system or filter can apply.[99]

---

[94] See https://newsroom.fb.com/news/2019/09/elections-and-political-speech/.

[95] See https://world-wide-what.tumblr.com/post/190101116282/the-internet-can-be-a-really-wonderful-place-its; https://about.fb.com/news/2020/06/coronavirus/, and https://blog.twitter.com/en_us/topics/company/2019/.

[96] See https://www.aspistrategist.org.au/twitter-data-shows-china-using-fake-accounts-to-spread-propaganda/.

[97] See https://www.blog.google/outreach-initiatives/google-news-initiative/elevating-quality-journalism/, https://www.facebook.com/journalismproject/coronavirus-update-news-industry-support, and https://newsinitiative.withgoogle.com/.

[98] See https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud, and WHO 'Stop The Spread' https://www.who.int/news-room/feature-stories/detail/countering-misinformation-about-covid-19'.

[99] See https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79.

This ecosystem includes the products explained below.

Search engines consist of software systems designed to search for information on the World Wide Web. They operate in an automated fashion using sophisticated algorithms to collect information, in a process known as 'crawling'. Web crawlers, commonly referred to as search engine bots or spiders, generally return results in a curated, ranked set of links to content websites.[100] Examples include Google Search, Baidu and Bing. This excludes downstream partners that host search functions on their own platforms that are powered by third-party search engines, as they have no legal or operational control of search results nor the order in which they are produced.

Software as a service (SaaS) allows users to licence software, often on a subscription basis, which is centrally hosted by a company. The infrastructure and data are hosted in the service provider's data centre, usually using cloud-based computing. End users control their usage of the software, and service providers have limited control over its usage once licensed. SaaS has become a common model for many applications, including office software, some messaging software, enterprise and creative tools. Some examples include Adobe's creative, marketing and business software, and Microsoft Office.

User-generated content platforms are online services that host high volumes of content uploaded by end users and enable them to connect with each other. User-generated content is often accessible in distinct links, and collections of content can be displayed in 'feeds' and curated by algorithms or displayed chronologically. User-generated content platforms can include blogs and microblogs, social media, social networks, discussion boards and photo, text and video sharing sites and some marketplaces. Examples include Facebook, Twitter, YouTube, LinkedIn and Change.org.

Messaging services and email permit sharing within 'closed' messaging groups or between individuals. These services may be products offered within or associated with user-generated content platforms (e.g. Facebook Messenger), or may also be offered with SaaS offerings (e.g. Slack, Outlook) or associated with particular hardware (e.g. Apple iMessage).

Digital content aggregation platforms are intermediaries that collect information from various sources and deliver content to consumers in a curated and branded news or information product. Users are generally able to filter and utilise custom tools to tailor the aggregated results to personal interests. Examples also include Google News, Apple News, and Flipboard. Aggregation technology is also central to some e-commerce and marketplace platforms (such as Redbubble). Whereas, other services which started as aggregators (Yahoo) have shifted to feature greater emphasis on original news content. These platforms either include third-party content, original content produced by the platform, or a combination of the two.

Industry initiatives: five themes

The following table outlines an industry framework of five common themes identified in the various efforts to counter mis- and disinformation. There is a spectrum of initiatives aimed at both mitigating and addressing mis and disinformation content and behaviour, empowering users of services with information, elevating quality content, and promoting digital media

[100] See https://www.accc.gov.au/system/files/ACCC%20Digital%20Platforms%20Inquiry%20-%20Preliminary%20Report.pdf p 23.

literacy.

Policies to respond to mis- and disinformation content

These are active measures to identify and address disinformation and harmful misinformation presently in place in content and platform policies, such as restrictions, community guidelines or terms and conditions that industry apply to users across their services and enforce through a range of proactive and reactive reporting mechanisms that include technical measures and human review. These generally focus on content restrictions for user-generated and advertiser content.

Measures to address inauthentic behaviour

These include efforts to address inauthentic behaviour, which is a key signal for disinformation in particular. This includes action on fake accounts, automated bots or other manipulative behaviour that are inauthentic or designed to deceive other users of the platforms.

Credibility signalling & contextual information

This encapsulates measures which are intended to assist users identify the reliability, trustworthiness and source of news content featured on a service. Our interviews with industry also indicated that these efforts often extend beyond news, in areas such as image authenticity. These initiatives can take the form of content 'badging', 'trust ticks', 'fact-check labels' or other forms of expandable information buttons that reveal the extent to which material has been verified, and collated with accountability, ethics, and the highest standards of practice. The surrounding information is intended to empower users with sufficient context to judge for themselves the accountability frameworks of a particular source.[101]

Measures to promote quality content

These are solutions which utilise a range of machine learning, algorithmic, human editorial, and curation processes to promote genuine and trusted content and information from trusted government or news sources, or high quality information that has been fact-checked, in order to improve the quality of content exposed to consumers.

Media literacy efforts to educate about mis- and disinformation

The recurring theme of digital media literacy recognises a healthy information ecosystem depends upon informed consumers of digital services, and increased media literacy is critical to empowering consumers in combatting information disorder.[102] Initiatives in this category also acknowledge the essential role which independent research has in identifying emerging issues and contributing to solutions that improve information outcomes for both industry and consumers.[103]

---

[101] Report from London School of Economics, Commission on Truth, Trust and Technology, 2019. See http://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis.pdf.

[102] Key Findings, *ACCC Digital Platforms Inquiry Final Report* June 2019, p.359; C Wardle & H Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policymaking, Council of Europe*, 2017.

[103] European Commission, 'Code of Practice on Disinformation One Year On: Online platforms submit self assessment reports', 2019.

TABLE 1: 'Five themes' – a current snapshot of industry initiatives, sampled from evolving policies and efforts across all platforms.[104]

	Policies to address to mis- and disinformation content	Measures to address inauthentic behaviour	Credibility signalling	Measures to promote quality content	Education and media literacy efforts
Microsoft	Microsoft's Code of Conduct covers the bulk of Microsoft's consumer products, websites, and services[105] and restricts fraudulent, false, or misleading behaviour and harmful activities**.** Microsoft Advertising features[106] also have disallowed product and services policies. LinkedIn's Professional Community Policies include measures to address harassing, hateful, violent and exploitative content, which it considers relevant in addressing some harmful misinformation and disinformation content. Violations are detected through mix of automated defences and user reports.[107] Search engine Bing has guidelines[108] that restrict inappropriate, manipulative, or misleading behaviour.	Microsoft successfully tested a new AI Framework in early fake news detection, called Multiple Sources of Weak Social Supervision (MWSS) reduces this timeframe. The new framework has tested successfully with user engagement on news articles. MWSS leverages weak social supervision signals from multiple sources, reducing aggregation times and making the approach more suitable for early detection.[109] LinkedIn has focussed on using automated, machine learned models to identify characteristics of bad actors/fake profiles.[110] LinkedIn utilises AI and machine learning in 'fake' account detection. 93% of blocked accounts for June-Dec 2019 were detected through automated measures.[111]	Microsoft uses the NewsGuard plugin tool for its Edge browser and Bing search engine. This tool generates trust certificates rating websites on nine journalistic standards criteria. Ratings are colour-coded and include green (pass), red (fail), yellow (satire).	LinkedIn has an editorial team of journalists globally including in Australia who work to create and curate information and promote it in various editorial products promoted to their members.	Microsoft has digital media literacy programs that utilise the tool NewsGuard. Public Libraries and schools across all markets are provided free access to the NewsGuard browser plug-in for use in their digital media literacy education programs. Microsoft also has a Defending Democracy program which increases political advertising transparency online, explores technological solutions to preserve and protect electoral processes, and defend against disinformation campaigns.

[104] Table 1 is provided as an indication of some industry initiates.

[105] See https://www.microsoft.com/en-ph/servicesagreement/#serviceslist.

[106] See https://about.ads.microsoft.com/en-au/resources/policies/disallowed-content-policies.

[107] See https://www.linkedin.com/help/linkedin/answer/34593/linkedin-professional-community-policies?src=li-other&veh=blog.linkedin.com%7Cli-other.

[108] 'Abuse and Examples of Things to Avoid', *Bing Webmaster Guidelines*. See https://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a.

[109] Shu et al, 2020 'Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News', April 2020. See https://arxiv.org/pdf/2004.01732.pdf.

[110] See https://engineering.linkedin.com/blog/2018/09/automated-fake-account-detection-at-linkedin.

[111] See https://about.linkedin.com/transparency/community-report#fake-accounts.

Twitter	In relation to coronavirus, Twitter broadened its definition of harm to address content that goes directly against guidance from authoritative sources of global and local public health information. It has also broadened its guidance on unverified claims related to COVID-19 that have the potential to incite people to action, could lead to the destruction or damage of critical infrastructure, or cause widespread panic or social unrest may be considered a violation of our policies Twitter has restrictions on synthetic or manipulated media that are likely to cause harm, including credibility signalling and transparency efforts in these areas. It also has relevant measures in areas user safety, privacy and authenticity, such as policies that apply to users who seek to manipulate trending topics lists and content that is considered likely to lead to imminent danger, harm, or violence.[112] It also has a range of relevant advertising policies, including the global prohibition of the promotion of political content and restrictions on state media purchase of advertising.	Twitter Community Rules contain a range of restrictions, including on 'platform manipulation', including to 'artificially amplify or suppress information' or 'engage in behaviour that manipulates or disrupts people's experience on Twitter'. It restricts impersonation, highlighting behaviour that may 'mislead, confuse, or deceive others'. Twitter uses machine learning, along with policies and human review, to determine how Tweets are presented in communal places like conversations and search. They detect for behaviours that distort and detract from the public conversation and use this to determine how Tweets are organised. This results in lower quality, unhealthy content becoming less visible and healthier, higher quality content more visible.[113]	Twitter introduced new labels and warning messages that will provide additional context and information on some Tweets containing disputed or misleading information related to COVID-19. These labels linked to a Twitter-curated page or external trusted source containing additional information on the claims made within the Tweet. The labels cover the content, requiring an extra click to view the original post.	Twitter collaborated with the Australian Department of Health, and other governments internationally, to develop a proactive prompt which directs users to authoritative information from the Government and WHO when people are searching for #COVID19 and related terms.	Twitter recently announced a global media literacy program with UNESCO. This partnership built on existing efforts where the two organisations have previously launched a media literacy focused handbook. Efforts in this area are focused on the verification of sources, critical thinking, active citizenship online, and the breaking down of digital divides. Twitter has other partnerships around journalism training and media literacy initiatives, include Reporters Without Borders, and the Reporters Committee for Freedom of the Press. These are aimed at ensuring Twitter's real-time capacity to neutralise misinformation is built into the newsroom approach of established media outlets. Twitter also maintains a public archive of state-backed information operations.

[112] Twitter announcement on a 'broadened definition of harm' to address content contrary to guidance from authoritative sources of global and local public health information. See https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#misleadinginformation

[113] See https://blog.twitter.com/en_us/topics/product/2018/Serving_Healthy_Conversation.html.

Facebook	Facebook has a range of relevant community and advertising standards. These include restrictions on 'Misinformation and unverifiable rumours that contribute to the risk of imminent violence or physical harm...'; '…Pages and domains that propagate misinformation'; Manipulated media that 'would likely mislead an average person to believe that a subject of the video said words that they did not say'. Its advertising policies restrict misinformation such that it 'prohibits ads that include claims debunked by third-party fact checkers or, in certain circumstances, claims debunked by organizations with particular expertise. Advertisers that repeatedly post information deemed to be false may have restrictions placed on their ability to advertise on Facebook.' Facebook's relevant policies in relation to coronavirus included limiting misinformation and harmful content, prohibiting exploitative tactics in ads, removing misinformation related to coronavirus on Instagram, and various advertising restrictions.	Facebook focuses on addressing what it calls 'coordinated inauthentic behaviour"\' which is where people or pages seek to mislead others (e.g. for financial or ideological purposes) about who they are or what they are doing . This approach focuses on deceptive behaviour, rather than content. Signals for this behaviour might include manipulation to make a network of accounts appear from one location when they are actually from another. This behaviour is detected and actioned by the platform through a combination of human investigators and technology that focus on the most sophisticated manipulation, as well as technology that proactively identifies patterns of this behaviour.[114]	Facebook News partners with third party fact-checking operations, via the non-partisan International Fact-Checking Network (IFCN). In Australia, third-party fact checking is provided by Agence France Presse and AAP. Items found to be 'false' are demoted in news feed. A 'Click Gap' signal then ensures better qualify content is prominently shared on its network and unverified content is de-emphasised on news feeds. Users attempting to share an item which has been factchecked as 'false' are prompted before doing so. Interstitial 'screens' are also used such that users have to 'click through' to access the content. In applying these measures as part of their coronavirus response, Facebook found that 95% of users did not 'click through' to view the 'false' content.[115] Facebook also includes user prompts such as 'Related Articles', 'Context Button', 'More from this publisher' and 'Shared by Friends' to display third-party fact checked articles.	In relation to coronavirus, Facebook introduced prompts in the Facebook News Feed and Instagram Feed of every Australian user, directing them to official Australian Government information, and had similar partnerships in other countries. They worked with the Australian Government and other partners such as Atlassian to release a chatbot on WhatsApp where Australians can access the latest Government coronavirus information. A Coronavirus Information Center on Facebook, was established to assist users to connect with authoritative information and resources via Facebook and Instagram, supported local news organizations, and the WHO Health Alert on WhatsApp.	Facebook has a worldwide journalism project providing news integrity initiatives to advance media literacy and increase trust in journalism.[116] It has a user-facing Digital Literacy Library, with education modules for young people in a range of areas including verification skills. Facebook publishes monthly reports available about 'coordinated inauthentic behaviour' takedowns which can be used by researchers wanting to understand disinformation.[117] Facebook announced $1m grants[118] to support coronavirus fact-checking. This is in addition to Facebook's existing Journalism Project which provides funding for training and resources to over 400 newsrooms worldwide. Further news resourcing initiatives in 2020 include an additional $125m to support the news industry coronavirus response. This includes grant resourcing for local news as part of the Facebook Journalism Project.[119]

[114] See https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

[115] See https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

[116] https://www.facebook.com/journalismproject.

[117] Facebook CIB reports, see https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.

[118] See https://about.fb.com/news/2020/07/coronavirus/#supporting-fact-checkers.

[119] See https://about.fb.com/news/2020/07/coronavirus/#news-industry-investment.

Google	Google enforces policies to address malicious behaviours and certain types of harmful misinformation. Policies across Google Search, Google News, YouTube, and advertising products outline behaviours that are prohibited– such as misrepresentation of one's ownership or primary purpose on Google News and advertising products, or impersonation of other channels or individuals on YouTube. In addition, policies also prohibit certain types of harmful misinformation: for instance, YouTube and Ads policies prohibit deceptive manipulated media or information about voting procedure or candidate eligibility that contradict official government records. Google advertising policies include a 'sensitive events' policy which prohibits advertising that may try to capitalise on tragic events such as a natural disaster, conflict or death. For example, under this policy, Google has blocked numerous ads attempting to capitalise on the coronavirus pandemic.	Google operates and enforces policies across its products such as Google Search, Google News, YouTube, advertising products that outline prohibited behaviours – such as misrepresentation of someone's ownership or primary purpose on Google News and advertising products, or impersonation of other channels or individuals on YouTube. Google Search actively looks for and targets attempts to deceive its ranking systems. Google News has restrictions on the impersonation of any person or organisation, sites or accounts that engage in coordinated activity to mislead users – including, but not limited to, sites or accounts that misrepresent or conceal their country of origin or that direct content at users in another country under false pretences. YouTube manages information tied to elections through effective ranking algorithms, and policies against users that misrepresent themselves or who engage in other deceptive practices.[120] Google communicates findings on government-backed phishing, threats and disinformation. The Google Threat Analysis Group has recently launched a new quarterly bulletin to share information about actions against accounts attributed to coordinated influence campaigns.[121]	Google provides users with sources of information and various safe navigation tools across its range of products. For example, Google Search and YouTube, display information panels in Search results to provide context and basic information about people, places, and events in relation to particular searches. Eligible channels on YouTube can apply for a verification mark which signals a channel is authentic - representing the real creator, brand, or entity it claims to be. Fact-check tags or snippets might show below links in Google Search and Google News, outlining that a specific piece of content purports to fact-check a claim made by a third party. For web developers building a web page that reviews a claim made by others, they can include 'ClaimReview' structured data on their web page in order to have a summarized version of the fact check to display in Google Search.[122]	Google products are equipped with tools to manage the vast amounts of material available on the web and deliver content tailored to users. For example, Google Search, Google News and YouTube utilise machine-based learning to elevate authoritative, high-quality information algorithms, apply non-partisan determination of news and search ranking and focus objectively on signals to detect inauthentic content.[123] YouTube also has specific product features to highlight authoritative content in the moments surrounding fast-developing breaking news events. These features included text-based information panels with information from news organisations and a link directly to the news website. YouTube also works with news producers to highlight breaking news video content on the YouTube homepage and in the YouTube, search results where users are displaying particular interest in a relevant topic. In relation to coronavirus, Google launched a COVID-19 microsite[124] featuring the latest official health updates and Google resources on various aspects of the pandemic, along with data and insights.	Google has developed and supported a number of programs to help users identify and avoid bad actors, and to better engage and make use of productive digital technology for the purpose of information discovery, communication and engaging with digital marketplaces. For example, the Google News Initiative (GNI) is a commitment of $300 million over three years to strengthen and elevate quality journalism on the web, including through building audience understanding and piloting digital publishing models. Digital Springboard is a free, in-person, digital skills training program offered through a national network of community organisations and institutions that promotes the core digital skills needed to thrive in work and life. The eSmart Digital License program is designed to help Australian children to play safe and stay safe, online. The program offers three different versions of age-appropriate content to help children understand what to look out for on the web, and how to deal with any threats when they arise. The Alannah and Madeline Media Literacy Lab - announced in 2019 and launched in July 2020 - is designed to teach students to critically analyse and navigate the online environment. It provides secondary school teachers with

[120] See https://kstatic.googleusercontent.com/files/ .

[121] See https://blog.google/threat-analysis-group/.

[122] See https://developers.google.com/search/docs/data-types/factcheck.

[123] See https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf p11.

[124] See https://www.google.com.au/covid19/.

					Australian curriculum-aligned content, classroom, and remote delivery ideas. Google also provides datasets and synthesised content for researchers working on AI detection tools.[125]
Apple	Apple's Podcast Connect policy covers submissions to Apple Podcasts, including Apple Podcasts for iOS and Apple Podcasts for Mac. These guidelines address inauthentic content 'designed to mislead' users and other overlapping policy areas to disinformation such as hate speech ('Nazi propaganda') and spam.[126]	Apple maintains policies which address elements of behaviours and content of mis- and disinformation in their App store and Podcast content.[127] Apple website terms include prohibitions on product enabled with automatic devices or other' or 'inauthentic' capabilities.[128]	Apple News is focussed on elevating the visibility of stories from credible, known news outlets. They also have Australian editorial team that curate and highlight high quality news content.[129]	In response to coronavirus, Apple introduced new App Store measures for App submissions. Apps that contain medical information must now be submitted by a recognised authority.	In 2019, Apple announced[130] a new literacy education program in conjunction with The News Literacy Project which offers nonpartisan, independent media literacy programs, including a 'misinformation guide'.

[125] See https://www.blog.google/outreachinitiatives/google-news-initiative/advancing-research-fake-audio-detection/.

[126] See https://help.apple.com/itc/podcasts_connect/#/itc1723472cb.

[127] See https://help.apple.com/itc/podcasts_connect/#/itc1723472cb.

[128] See https://www.apple.com/legal/internet-services/terms/site.html.

[129] See https://appleinsider.com/articles/19/05/12/editorial-can-apple-news-kill-fake-news-and-save-journalism.

[130] See https://www.apple.com/newsroom/2019/03/apple-teams-with-media-literacy-programs-in-the-us-and-europe/.

Other companies	Redbubble Community and Content guidelines include the prohibition of 'harmful misinformation' which it defines as 'any misleading or false information that harms or significantly threatens public safety'.[131] Change.org has restrictions in its Community Guidelines on misleading content and will 'remove content which is verifiably incorrect and which has the potential to cause harm to our users'. It also has restrictions on impersonation and hate speech, which it considers relevant in addressing some harmful misinformation and disinformation content.	Verizon Media has a policy on non-genuine behaviour, as well as restrictions on content designed that could mislead, defraud, or otherwise, this includes attempts to disenfranchise voters or otherwise maliciously interfere in elections. This includes restrictions on causing 'confusion between you and any other person, organization, or company, or mislead users about the origin of the content you post or your affiliation with any other person, organization, or company.'	Adobe is working with software tool companies, publishers, social media companies, human rights organisations and academic researchers to develop an open industry standard for content attribution. Creators and publishers will be able to imprint data attribution on material they create and share. As a result, users (individuals, or news organisations) will be able to determine the provenance of an item using the metadata and determine whether it has been manipulated.[132] .	Verizon Media created a coronavirus hub, across the Yahoo ecosystem that includes real-time news about the global pandemic.[133]	Adobe and UC Berkeley researchers have collaborated on AI research designed to detect modification of images made with Photoshop's Face Aware Liquify feature.[134] The Trust Project (an international consortium of news operators) has developed 'trust indicators' which are used to surface and display quality journalism across a wide range of platforms[135]

[131] See https://help.redbubble.com/hc/en-us/articles/202270929-Community-and-Content-Guidelines#misinformation.

[132] See https://theblog.adobe.com/adobe-reinforces-commitment-to-content-authenticity-previews-technical-white-paper/.

[133] See https://www.verizon.com/about/news/our-response-coronavirus.

[134] See https://theblog.adobe.com/adobe-research-and-uc-berkeley-detecting-facial-manipulations-in-adobe-photoshop/.

[135] See https://thetrustproject.org/trust-project-launches-indicators/.

Industry initiatives in depth

Here we further illustrate the details of some of the initiatives in Table 1, under the five common themes identified above.

Policies responding to mis- and disinformation

In February 2020, Twitter implemented a policy to manage inauthentic content, in addition to its efforts focused on platform manipulation behaviour.[136] The new rules prohibit users from deceptively sharing synthetic or manipulated media that are likely to cause harm. This included the initiative to label Tweets containing synthetic and manipulated content in order to better understand the context of information. The new approach uses the following criteria:[137]

1. Whether media is synthetic or manipulated and the degree to which it has been edited to alter composition (i.e. sequence, timing, or framing) and whether the affected media contains a real person in a fabricated or simulated circumstance.
2. Whether the media was shared in a deceptive manner and the motivation of the sharer. This assessment involves considering the context of surrounding material, associated tweets, meta data and the profile of person, and questions around whether the content likely to impact public safety or cause serious harm.

Tweets most likely to be removed are those sharing manipulated media which are likely to cause harm. Considerations include: threats to physical safety of other people; risks of mass violence or civil unrest; targeting others with aim to silence; or threatening privacy or ability of others to freely express themselves.

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

Twitter's action based on three categories of manipulated media.[138]

Measures to address inauthentic behaviour

Facebook's approach to disinformation is focused on addressing 'coordinated inauthentic behaviour' (CIB) that seek to manipulate the public. As well as the use of technology, Facebook has a cross-disciplinary team of over 200 people focused on finding and disrupting the following aspects.

[136] See https://help.twitter.com/en/rules-and-policies/manipulated-media.

[137] See https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html.

[138] See https://help.twitter.com/en/rules-and-policies/manipulated-media.

1. Sophisticated influence operations aimed to manipulate public opinion. It considers these activities to be largely politically motivated, aimed at gaining influence for a strategic goal. Within this category, Facebook observes two further categories activities that it works to stop
    - CIB in the context of domestic, non-state campaigns. When this is discovered, Facebook removes both inauthentic and authentic accounts, Pages and Groups directly involved in this activity.
    - CIB from foreign or government actors. When this is discovered, Facebook employs broad enforcement measures including the removal of every on-platform property connected to the operation itself and the people and organisations behind it.
2. High volume inauthentic behaviours like spam and fake engagement. It considers these activities to be largely financially motivated.

In order to maintain what it calls 'continuous enforcement', Facebook uses automated and manual detection to remove accounts and Pages connected to networks previously removed.

For the last three years, Facebook has released periodic reports on this behaviour for law enforcement, researchers and the public to better understand the nature of this manipulation, with a focus on the first category above of sophisticated influence operations. In April 2020, Facebook removed eight networks of accounts, including two foreign or government actors from Russia and Iran, and six domestic operations within the US, Georgia, Myanmar and Mauritania.[139]

Credibility signalling

Adobe is working to establish open industry standards for content authentication for digital media. The Content Authenticity Initiative (CAI) is an initiative announced by Adobe in November 2019[140] in partnership with Twitter and the New York Times. The project is aimed at attribution in the creation of digital content, such as images.

The initiative recognises that content attribution for creators and publishers is essential for user trust yet balances this with the fact that modification is often a necessary part of creative process. That is to say, not all 'altered' content is mis- or disinformation; instead, CAI aims to balance that challenge by providing users with information to discern for themselves what is malicious.

Central to this idea is development of an open industry standard with cross-industry participation designed to detect and communicate to users, the provenance of a modified item of digital content ('asset'). This will allow end users to evaluate an 'asset' and discern for themselves whether content is mis- or disinformation, within the particular context its being viewed. This information may consist of where and when a picture was taken and by whom. If speech or voice manipulation is part of that assessment, a user may be provided with information on the video's voice speed compared with the technical standard for that type of content.[141]

A summit in early 2020 brought together stakeholders from technical and content teams, to launch collaborative working groups and partnerships on designing an attribution tool to assess content

[139] See https://about.fb.com/news/2020/05/april-cib-report/.

[140] CAI Announcement from Adobe. See https://s23.q4cdn.com/979560357/files/doc_events/2019/11/1/110419AdobeNYTandTwitterAnnounceContentAuthenticityInitiative.pdf.

[141] See https://theblog.adobe.com/adobe-reinforces-commitment-to-content-authenticity-previews-technical-white-paper/

authenticity and the provenance of digital content.[142] The discussion to date has considered three key areas:

1. **Detection** of 'deep fakes'[143] and similar content needs a refined approach. Algorithms and manual detection are able to identify intentionally misleading, but these must keep pace with the increasing sophistication of editing tools to remain effective. Detection must also balance the fact that not all manipulated content is malicious; for example, movies are edited, photographs are enhanced for aesthetics.
2. **Attribution** or version history can empower users with information about who created, altered and shared a particular piece of media. However, care needs to be taken in using the solution so as to mitigate unintended consequences. For example, care must be taken so as not to invalidate or create risk for genuine photojournalists who may be reliant on anonymity to carry out their work.
3. **Consumer education** will assist creators to understand disinformation and work with tools and techniques to eliminate it. Programs focused on this skill can equip consumers with tools and information to better evaluate digital media and understand it with a more discerning view.

The CAI collaborators published a whitepaper *Setting the Standard for Content Attribution*[144] in August 2020 in which they stated:

> The initial mission of the CAI is to develop the industry standard for content attribution. We will provide a layer of robust, tamper-evident attribution and history data built upon XMP, Schema.org and other metadata standards that goes far beyond common uses today. This attribution information will be bound to the assets it describes, which will in turn reduce friction for creators sharing the attribution data and enable intuitive experiences for consumers who use the information to help them decide what to trust.

The whitepaper also went on to highlight that

> Increasing trust in media requires the ongoing engagement of diverse communities. The CAI does not prescribe a unified single platform for authenticity, but instead presents a set of standards that can be used to create and reveal attribution and history for images, documents, time-based media (video, audio) and streaming content. Although the initial implementations will focus on imagery, the initiative aims to specify a largely uniform method for enabling attribution from various points of view through which diverse stakeholders can build decentralized knowledge graphs about the trustworthiness of media.

Google recently announced measures to address authenticity in image search, particularly images recirculating in viral misinformation cycles. Fact-check labelling will be utilised to provide context for image search results. The initiative will draw on services provided by third-party fact-checkers and publishers who will now be able to tag fact-checked images using ClaimReview (a method for publishers to communicate to search engines that an image has been verified, described in Table 1).

Other initiatives reach beyond the realm of images. Facebook has been working in partnership with

---

[142] See https://theblog.adobe.com/the-content-authenticity-initiative-summit-collaborating-to-drive-trust-and-transparency-online/.

[143] The term 'deep fakes' is defined as realistic photo, audio, video, and other forgeries generated with artificial intelligence (AI) technologies. See Kelly M. Sayler & Laurie A. Harris, CRS In Focus IF11333, *Deep Fakes and National Security* (Oct. 14, 2019).

[144] See https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A2c6361d5-b8da-4aca-89bd-1ed66cd22d19#pageNum=1.

IFCN organisations on methods to identify and constrain the sharing of false news. Technology and machine learning are utilised to identify potential material likely to contain misinformation and prioritise those for third party fact-checkers to review and rate. The ratings which can be applied are 'False', 'Partly False', 'False Headline' or 'True'. Generally, a false rating will result in Facebook lowering the article in its News Feed with the aim of reducing exposure to the false news. The News Feed also displays articles on the same topic, from third-party fact-checkers, immediately below the false story.

Microsoft partners with NewsGuard, using its credibility ratings plugin on its Edge browser and Bing search engine. NewsGuard provides 'nutrition' labelling on predominately 'hard news' sites (i.e. not e-commerce, 'entertainment' or sports sites, etc). Each label provides editorial information assessed against nine criteria that NewsGuard has developed. These include Red and Green (pass/fail), Yellow (satire) and Grey (platform/user generated information). The aim is for transparency, and users can also assess information based on the criterion of individual importance.

Twitter also introduced fact-checking and credibility measures to counter misleading information. These have been appearing as labels and warning messages attached to a fact-checked tweet and which provide additional context or alternative sources of information on Tweets containing disputed or misleading information. Depending on the nature of the information, Twitter may decide to action in the following ways:

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

Twitter's action based on three broad categories [145]

Measures to promote quality content

In its News and Search services, Google has elevated high quality information in spaces such as 'Top Stories' Carousel or our 'News' Tab. A similar method applies to YouTube content, which highlights relevant and verified news content on its homepage. In the context of its coronavirus response, Google enhanced their search so searches for virus information prompted an 'SOS Alert' which returned prominently displayed news and information from trusted health sources including the WHO and Centre for Disease Control.

LinkedIn has focussed on providing accurate public health information during the coronavirus pandemic through editorial curation. 'Daily Rundown' is LinkedIn's editorial function developed that utilises push notifications to distribute health and economic recovery information from authoritative and verified sources to its members. Their editorial team also curates news stories with surrounding context from other verified sources (e.g., government policy announcements, public event) and where relevant, may add high quality posts from LinkedIn users.

As detailed in Table 1, the coronavirus pandemic saw a range of digital services partner with the

[145] See https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

Australian Government and other health authorities to elevate alerts and other credible information to guide Australians in response.

Education and media literacy efforts

In addition to a focus on information, various education and media literacy initiatives have also focused on images. Facebook and Microsoft have invested in the Deepfake Detection Challenge research initiative. This is a program focussed on building better tools for deep fake detection and recognised the collaborative effort required to produce solutions. Technical researchers around the world are given the opportunity to access grants and data sets to develop innovative new technologies for use in managing inauthentic content, particularly manipulated media.

Facebook has also partnered with Reuters, the world's largest multimedia news provider, to help newsrooms worldwide to identify 'deep fakes' and manipulated media through a free online training course. This is one example of a broader trend in such efforts to initiate partnerships with news organisations, such as the Google News Initiative (GNI) which provides fellowships, training, technology research and grants for news organisations, journalists and fact checking outlets.

Industry also partner with civil society on educational efforts. For example, Twitter has partnered with UNESCO to publish a new handbook for educators, entitled Teaching and Learning with Twitter aimed at raising awareness of media and information literacy among parents, educators. Google has partnered with the Australian non-profit Alannah & Madeleine Foundation on the development digital media literacy tool, which teaches high school students to critically analyse and navigate the online environment.

Investment in this area also extends to support for research and institutions involved in undertaking vital research in machine learning and other non-technical disinformation responses. On this front, Twitter has released several tranches of data sets from European, Asian and Middle East jurisdictions, for independent analyses and research into platform manipulation across several jurisdictions.

Together, the main platforms are a key contributor to various First Draft News projects. Twitter, Facebook (via the Journalism Project) and Google's News Initiative each provides resources, and collaboration on global projects such as the CrossCheck initiative and First Draft's coronavirus resources hub for reporters.

3 International initiatives

The leading example of regulatory initiatives to address disinformation is the EU Code of Practice on Disinformation. Having been implemented in late 2018, it has been adopted by a number of global digital platforms and is now the subject of independent assessment. While we considered the definitions used under this Code and its key commitments in relation to the draft Australian Code of Practice in Section 1, in this section we explore the Code in further depth in relation to the scope of the commitments and its widespread adoption. We also explore approaches adopted in other jurisdictions to mis- and disinformation.

Our review of other jurisdictions has not revealed a country-level approach or instrument directly comparable to the EU Code. The closest case is the code developed in Taiwan, although there is also a code implemented in India designed to strengthen confidence in the election process.

The review of jurisdiction has, however, revealed a range of regulatory and non-regulatory measures to address disinformation. Proactive measures adopted by government, many being education-based and some in collaboration with industry, are attempting to either stop the spread of disinformation or to replace false information before it has time to spread.

Some initiatives involve government in detecting false information and, while innovative, there does appear to be some risk that measures of this kind could impede on freedom of expression by curbing commentary on government initiatives. There is a public debate over this aspect in South Korea, with some advocating greater use of self-regulation and others encouraging government intervention.

This initial review of other select jurisdictions revealed an interesting divergence in aspects that are considered important to address.

- A number of countries are attempting to strengthen confidence in the electoral process, but for others (such as the Czech Republic and South Korea) a primary concern is the presence of a neighbouring state known to be active in the spread of disinformation.
- For both India and Taiwan, disinformation about natural disasters is of particular concern. In both countries, legislation has been used for these measures; legislation is also being contemplated in the US to address the specific problem of deep fakes.
- In Canada, where, since May 2019 there has been a Digital Charter that includes disinformation, legislation has been used in a more general way to promote transparency, specifically recognising the role that digital platforms play in modern democracies, while encouraging platforms to enforce to policies to limit the potential that they are manipulated to spread disinformation.

We now turn to the EU Code and then several other international initiatives.

The EU Code of Practice

The Code and its signatories

The European Union Code of Practice on Disinformation ('the EU Code') is a voluntary, self-regulatory code that is designed to minimise the spread of online disinformation and fake news. The EU Code recognises the importance of open and transparent debates for democracy and broader civil society. As noted above, there are three principal elements in its definition of 'disinformation': 'verifiably false or misleading information' which 'is created, presented and disseminated for

economic gain or to intentionally deceive the public' and which 'may cause public harm'.[146]

To address the challenge posed by disinformation the Code sets out 11 objectives that signatories recognise as important in efforts to address the dissemination of disinformation. These objectives are followed by 15 commitments that signatories can choose to commit to. These commitments cover five areas:

1. Scrutiny of ad placements
2. Political advertising and issue-based advertising
3. Integrity of services
4. Empowering consumers
5. Empowering the research community.

Each signatory chooses its own commitments, allowing it to cater its response to the nature of the organisation.[147]

The Code also sets out reporting provisions. There are a further six commitments under measuring and monitoring the Code's effectiveness. Broadly, signatories are to provide self-assessments of their performance, evaluated against the commitments under the Code that they have entered. Initially, these reports were to be provided monthly, from January to May 2019, and aimed to coincide with the European elections. Following this, the first annual self-assessments were provided in October 2019. Based on these reports there has so far been:

- A summary and analysis of the self-assessments conducted by the EU Commission, and
- A report published by the European Regulators Group for Audiovisual Media Services.[148]

The Code adopts the definition used in the European Commission Communication 'Tackling online disinformation: a European approach'. As explored in Section 1, this defines disinformation as, 'verifiably false or misleading information' which, cumulatively,

(a) "is created, presented and disseminated for economic gain or to intentionally deceive the public"; and
(b) "may cause public harm", intended as "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security."'[149]

The Code clarifies what is *not* disinformation. Particularly, disinformation 'does not include misleading advertising, reporting errors, satire and parody, or clearly identified partisan news and commentary, and is without prejudice to binding legal obligations, self-regulatory advertising codes, and standards regarding misleading advertising.'[150]

In January 2020, the Code had 15 signatories. This covered the major platforms (Facebook, Google, and Twitter), tech companies (Microsoft and Mozilla), and trade associations and other

[146] As noted in Section One above, the definition includes within it a definition of 'public harm' and is followed by a note on scope which excludes content such as misleading advertising. See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.

[147] European Commission, 'EU Code of Practice on Disinformation' (26 September 2018). See https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation, ('EU Code'), p 1.

[148] See https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf..

[149] EU Code 1 referencing: 'European Commission Communication 'Tackling Online Disinformation: A European Approach' paragraph 2.1. In paragraph (b), 'intended as' refers to the original definition in the Communication, which put it this way: 'Public harm comprises threats to democratic political and policy-making processes as well as …'. See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.

[150] EU Code (n 1) p 1.

organisations.[151] In June 2020, TikTok became a signatory.

Code development

The threat of disinformation was first flagged by the European Council in 2015. Initially focusing on the issue of 'fake news' the approach of the Commission shifted to focus on disinformation in part in response to the Cambridge Analytica incident.[152] In April 2018, an EU-wide Code of Practice on Disinformation was first proposed. This Code was announced in September 2018 and signed in October 2018. This Code is accompanied with the broader European Action Plan against Disinformation.

The Commission plays an oversight role regarding the implementation of the Code, reporting back to the European Council. As mentioned above, the Commission was responsible for collating a summary and analysis of the initial monthly reports and the first annual reports provided by signatories. Further analysis of the functioning of the Code is conducted by external organisations under the direction of the Commission, with an assessment delivered in 2020.

The process also included a Sounding Board – a committee of representatives from media, civil society, fact checkers and academia established to provide an opinion on the drafting of the Code. This committee flagged two initial concerns: the absence of clearly measurable key performance indicators or other measurable objectives; and a perceived lack of clear and meaningful commitments forming common guidelines for signatories which was said to limit its effective operation as self-regulation).[153] However, the Commission disagreed with this comment, saying the Code is consistent with existing principles for self-regulation set by the Commission.[154] These criticisms are incorporated into commitments 16-21 of the Code which spell out Key Performance Indicators for reporting on the effectiveness of the Code.

Implementation and assessment

The reporting obligations under the EU Code have been extensive. Every month from January to May 2019, Facebook, Google and Twitter were obligated under the Commission's action plan to demonstrate how they are fulfilling the requirements of the Code.[155] They were then required to provide annual reports. The Commission would then provide its own assessment each month. It notes some reservations, particularly around metrics used and explanations of action taken, in the initial month.[156] In the final monthly report, the Commission noted improvements – for example, Google, Facebook and Twitter had all improved the scrutiny of ad placements to limit malicious click-baiting practices and reduced advertising revenues for spreaders by, for example, removing

[151] Links to the signatures can be accessed here: https://ec.europa.eu/digital-single-market/en/news/roadmaps-implement-code-practice-disinformation.

[152] Peter Chase, 'The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem' (29 August 2019), p 3.
Accessible at: <https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf>

[153] Sounding Board, 'The Sounding Board's Unanimous Final Opinion on the So-Called Code of Practice' (24 September 2018). See https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[154] Peter Chase, above, 9.

[155] European Commission 'Code of Practice against disinformation: Commission recognises platform's efforts ahead of the European elections'. See https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_19_2570.

[156] European Commission, 'Code of Practice on Disinformation Intermediate Targeted Monitoring – Intermediate Targeted Monitoring – January Reports'. See https://ec.europa.eu/digital-single-market/en/news/first-monthly-intermediate-results-eu-code-practice-against-disinformation.

ads and closing ad accounts as a result of deceptive or inauthentic behaviour.[157]

In its report on the first annual reports, the Commission notes, for example, higher transparency around platforms' policies addressing disinformation, efforts by platforms to disrupt advertising and monetisation connected with disinformation, and measures to increase transparency of political advertising. But it also noted continuing reservations about the metrics provided by the platforms and a lack of progress on joined-up efforts 'to identify persistent or egregious purveyors of disinformation and develop indicators for the trustworthiness of media sources, for the development and deployment of ad scrutiny and brand safety measures'.[158]

Critiques of the EU Code

An independent assessment of the EU Code was conducted by The European Regulators Group for Audio-visual Media Services (ERGA).[159] ERGA regarded the Code as an important step in the process of building a new relationship between its signatories, the EU and National AV Regulators. However, it considers there is a need for greater transparency about how the Code is being implemented, noting also that the self-reporting cannot be independently verified and that there is also no uniformity in the procedures and the definitions that have been adopted by the different platforms. ERGA suggests that all of the platforms be required to comply with the same obligations in a uniform manner and adopt more precise definitions, procedures and commitments. Paul-Jasper Dittrich argues for an EU statutory layer of general principles, a co-regulatory layer comprising an industry-developed Code, and company-specific measures to implement the Code which are approved by the EC.[160] Separately, James Pamment[161] has noted that although there have been areas of progress, the weak points of this system include how there is a lack of detail of data in the signatories' reports and success metrics for their efforts, and an inconsistency of approaches. He states that the inconsistent terminology 'indicates a lack of consensus among key stakeholders regarding the scope of the issue and therefore its potential solutions'. He states that clarity over objectives and terminology is required.

These criticisms appear consistent with comments in follow-up to recent first phase baseline reporting which formed part of the Code's monitoring and reporting programme.[162] Despite overall

[157] European Commission, 'Code of Practice on Disinformation Intermediate Targeted Monitoring – Intermediate Targeted Monitoring – May Reports'. See https://ec.europa.eu/digital-single-market/en/news/last-intermediate-results-eu-code-practice-against-disinformation.

[158] European Commission, 'Annual self-assessment reports of signatories to the code of practice on disinformation 2019', (29 October 2019). See https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019 p 5, 9, 12.

[159] ERGA 'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice'. See https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.

[160] Paul-Jasper Dittrich: 'Tackling the spread of disinformation Why a co-regulatory approach is the right way forward for the EU' (December, 2019) *Jacques Delors Centre: Hertie School*. See https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/EZ_JDI_BST_Policy_Paper_Disinformation_Dittrich_2019_ENG.pdf p 7.

[161] James Pamment, 'EU Code of Practice on Disinformation: Briefing Note for the New European Commission' (March, 2020). See https://carnegieendowment.org/2020/03/03/eu-code-of-practice-on-disinformation-briefing-note-for-new-european-commission-pub-81187.

[162] 'First Baseline Reports – Fighting COVID-19 disinformation Monitoring Programme', European Commission (Web Page, 10 September 2020). See https://ec.europa.eu/digital-single-market/en/news/first-baseline-reports-fighting-covid-19-disinformation-monitoring-programme.

praise for signatories' progress on implementing policies cross the five pillars of the Code,[163] the Commission's assessment expressed shortcomings in the 'lack of common understandings of the scope of fundamental concepts and of uniform definitions of key operational terms [which] inhibits the effective implementation of measures by the signatories'.[164] Of particular note were the lack of uniformity of reporting procedures, fact-checking approaches and distinctions between types of false or misleading content and manipulative behaviour intended to amplify its dissemination online. These were considered 'necessary for framing appropriate responses by the platforms and other relevant stakeholders'.[165]

Such critiques need to be balanced with goals in Australia to have a code adopted by a variety of digital services that will have arguably different approaches, and varying capabilities for reporting.

India

For the 2019 general elections, the social media platforms (including Facebook, WhatsApp, Twitter, Google, ShareChat, TikTok) and an industry body, The Internet and Mobile Association of India, agreed to a Voluntary Code of Ethics which was in effect from 20 March 2019 until the general elections.[166] This was at the request of the Election Commission of India, which summoned these organisations to introduce a voluntary code of ethics.[167] The purpose of the Code was to 'identify the measures that Participants can put in place to increase confidence in the electoral process'. In September 2019, it was announced that the social media platforms, as directed by the Electoral Commission, have agreed to follow this voluntary Code of Ethics for all future elections.[168]

It is a three-page document where eight commitments are listed. It does not refer to 'disinformation' or related terms; instead, participants made general commitments, such as to 'facilitate access to information regarding electoral matters' and 'to voluntarily undertake information, education and communication campaigns to build awareness including electoral laws and other related instructions'.

In late 2019, the Press Information Bureau (a government agency) set up a fact-checking unit to verify news that relates to the Indian government.[169] The West Bengal government has also been preparing a database of fake news stories that have been distributed on social media over the past

[163] The Staff Working Document 'sets out the key findings of the Commission services' assessment of the implementation and effectiveness of the Code of Practice on Disinformation during its initial 12-months period of operation'. See https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212.

[164] European Commission, *'Staff Working Document: Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement'*, SWD (2020) 180 (10 September 2020) p 13.

[165] Ibid p 12.

[166] Tariq Ahmed 'Government Responses to Disinformation on Social Media Platforms: India' (September, 2019) *Library of Congress* (September 2019). See https://www.loc.gov/law/help/social-media-disinformation/india.php.

[167] Yatti Soni 'Ahead of State Elections, Social Media to Follow Voluntary Code of Ethics' (September, 2019) *INC 42.* See https://inc42.com/buzz/ahead-of-state-elections-social-media-to-follow-voluntary-code-of-ethics/.

[168] Taruka Srivastav 'Social Media Platforms Agree to Follow 'Code of Ethics' In India for Elections' (September, 2019) *The Drum*. See https://www.thedrum.com/news/2019/09/27/social-media-platforms-agree-follow-code-ethics-india-elections.

[169] 'Press Information Bureau sets up fact-checking unit to combat fake news related to govt' (November,2019) *The Print*. See https://theprint.in/india/press-information-bureau-sets-up-fact-checking-unit-to-combat-fake-news-related-to-govt/328248/.

few years.[170] Furthermore, the Uttar Pradesh police have set up 'digital armies', which comprise of prominent residents along with ex-service personnel, teachers, doctors, advocates, and journalists.[171] This involves WhatsApp groups being formed to keep an eye on potential disinformation and other damaging posts.[172] All the state police stations will manage these groups and the 'digital volunteers' will share posts that spread disinformation with the police as well as disseminate correct information.[173]

In terms of law, there is no specific provision in Indian law that deals with fake news.[174] However, there are several offences in India's Penal Code that criminalise certain forms of speech that may be able to be invoked in cases of misinformation.[175] Moreover, there are other relevant laws. For example, according to the Disaster Management Act, it is a crime to make or circulate a false alarm about a disaster or its severity. Furthermore, internet shutdowns by the Indian government are not uncommon.[176] In October 2018, it was reported that the Indian government 'turned off' the internet more than 100 times in 2018 to curb the spread of rumours on WhatsApp.[177]

Sweden

The Swedish Civil Contingencies Agency (knows as the MSB) is a government agency with the task of increasing awareness among residents of the threats that arise with disinformation and influence campaigns.[178] The MSB updated its public emergency preparedness brochure so that it has a section on disinformation.[179] In the lead up to the 2018 September elections, MSB educated local election authorities and various governmental bodies on how to detect influence campaigns from foreign entities.[180] Furthermore, it published a handbook in 2018 called *Countering Information Influence Activities: A Handbook for Communicators* which provided resources for people working in public administration.[181] Interestingly, the 2018 Minister for Digitisation worked alongside Facebook to establish a Facebook 'hotline'[182] where both the MSB and all political parties could let Facebook know if they came across problematic content during the election campaign.

Through the Swedish Innovation Authority, the government has also invested in a 'new digital

[170] Daniel Funke and Daniela Flamini 'A Guide to Anti-Misinformation Actions Around the World' *Poynter*. See https://www.poynter.org/ifcn/anti-misinformation-actions.

[171] Ibid.

[172] 'UP Police's 'Digital Armies' to curb fake news on social media'(July, 2018) *Economic Times*. See https://economictimes.indiatimes.com/news/politics-and-nation/up-polices-digital-armies-to-curb-fake-news-on-social-media/articleshow/65091621.cms?from=mdr.

[173] Ibid.

[174] Above n 166

[175] Ibid.

[176] Funke and Flamini 'A Guide to Anti-Misinformation Actions Around the World'.

[177] Timothy Mclaughlin 'How WhatsApp Fuels Fake News and Violence in India' (December, 2018) *Wired.* See https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/.

[178] Elin Hofverberg, 'Government Responses to Disinformation on Social Media Platforms: Sweden' (September,2019) *Library of Congress.* See https://www.loc.gov/law/help/social-media-disinformation/sweden.php.

[179] Ibid.

[180] Christina La Cour, 'Governments Countering Disinformation: The Case of Sweden' *Disinfo Portal*. See https://disinfoportal.org/governments-countering-disinformation-the-case-of-sweden.

[181] Hofverberg, 'Government Responses to Disinformation on Social Media Platforms: Sweden'

[182] Ibid.

platform' that is designed to curb the spread of online disinformation.[183] It has been funded by 'Swedish Television' and other Swedish broadcasters and this platform has three functions to help filter news: 'an "automated news assessment service" for evaluating news, a "personalised engine" for countering filter bubbles and a "fact assistant" for automating fact-checking processes and discarding fake and irrelevant news'.[184] Additionally, the State Media Council (a government agency) has developed teaching materials to help students learn to identify online disinformation.

Foreign law expert Elin Hofverberg writes that Sweden has criminalised many acts that relate to the dissemination of propaganda.[185] For example, accepting remuneration from a foreign entity to spread propaganda in Sweden is a crime. Additionally, spreading information that could be dangerous to the national security of Sweden is a crime. She also writes that it is a crime 'to intentionally affect public opinion or limit the freedom of a political organisation or a union or trade association to act and thereby jeopardize the freedom of speech and association through the use of force, coercion, or criminal threats'.[186] Additionally, spreading information that could be dangerous to the national security of Sweden is a crime, as is accepting some form of remuneration from foreign entities to spread disinformation in Sweden.[187]

Taiwan

Digital platforms have collaborated on a code of practice in Taiwan, a territory that faces particular problems with disinformation.[188]

The Code begins with a preface explaining that the guidelines allow the participating parties to adopt various approaches when implementing the guidelines. The Code then outlines its goal which is 'to unite non-governmental forces in Taiwan so as to promote the prevention and control mechanism of disinformation'. Next, the Code provides a definition for disinformation.

> 'The term "Disinformation" referred to in the Guidelines should conform to all the following three descriptions. When determining whether a piece of information can be classified as "disinformation", one should strictly abide by the freedom of speech and take into account international academic and practical consensus:
>
> 1) For the purpose of maliciously deceiving the public or creating improper economic gains (malicious-intention)
> 2) So as to create and spread verifiable false information or misleading information (false-action)
> 3) And is possibly compromising the sound operation of democratic politics or public safety (harmful-result)

[183] Above n 180.

[184] Ibid.

[185] Elin Hofverberg, 'Government Responses to Disinformation on Social Media Platforms: Sweden' (September,2019) *Library of Congress.* See https://www.loc.gov/law/help/social-media-disinformation/sweden.php.

[186] Above n 181.

[187] Ibid.

[188] The Code is publicly available only in Mandarin and has been translated by a NAATI-credentialled translator for this analysis. The title has been translated as 'Self-discipline Practice Guidelines for Disinformation Prevention and Control', although a Mandarin-speaking legal academic consulted on this issue advised that 'self-discipline' can also be translated as 'self-regulation'. For the Mandarin version, see https://www.tahr.org.tw/sites/default/files/u87/190621_disinformation_code_of_practice_taiwan.pdf.

The concept of "disinformation" in the Guidelines does not include wrong reports, satirical and imitative works and commercial advertisements that are not politically misleading.'

The Code then goes into the four points, which form the 'Content of the Guidelines'. These four points are the participating parties' commitment to:

- 'continuous investment in technology for the purpose of establishing a disinformation prevention mechanism and relevant safeguards';
- 'continuous increase in advertisement transparency and management';
- 'cooperating with a third party and government authorities to establish and maintain an independent, transparent and impartial supervision mechanism'; and
- 'through the training of digital literacy and media literacy, assist the public in acquiring the ability to identify disinformation'.

There are three or four more specific commitments under each of these areas.

Finally, the Code ends with the 'Implementation and Prospect' section where it states that the participating parties implement part or all of the guidelines, agree to conduct regular reviews and 'continue to have organisational conversations with relevant government authorities in an active manner.'

Separately from the operation of the Code, the government works in cooperation with civil society actors and fact-checking groups outside government. Two non-profit organisations, *Taiwan Media Watch* and the *Association for Quality Journalism*, have jointly founded the 'Taiwan Fact Check Centre'.[189] This fact check centre uses a 'back-end tool' that is provided by Facebook to track viral posts that are misleading and will fact-check them.[190] Once the post is confirmed to be incorrect, Facebook will inform anyone who had shared the post that it was not true.[191] The government of Taiwan as well as political figures and had encouraged people to engage with the FactCheck centre.[192]

In addition, the Taiwanese government requires state agencies to refute false claims that relate to their areas of responsibility on social media and the Internet within two hours.[193] This negation must be communicated in 200 characters or less, and in two different ways e.g., a picture, a short text, a video.[194] The reason for this is so that the rebuttals go viral before the fake news reaches an audience. Flemming Rose writes that 'humour is also an important element in the government's strategy, countering and minimizing manufactured outrage'.[195] Also, the government has introduced a new curriculum in school that focuses on media competence.

---

[189] Keoni Everington, 'Taiwan's first fact-checking center launches to battle fake news' (August 2019) *Taiwan News.* See https://www.taiwannews.com.tw/en/news/3497482.

[190] Emily Feng 'Taiwan Gets Tough On Disinformation Suspected From China Ahead Of Elections (December,2019) *NPR.* See https://www.npr.org/2019/12/06/784191852/taiwan-gets-tough-on-disinformation-suspected-from-china-ahead-of-elections.

[191] Ibid.

[192] Ralph Jennings 'Which Coronavirus Report are Fake? Ask These Fact Checkers' (February 2020) *Voa News.* See https://www.voanews.com/science-health/coronavirus-outbreak/which-coronavirus-reports-are-fake-ask-these-fact-checkers.

[193] Fleming Rose 'The Taiwan Election: Dealing with Disinformation while Protecting Speech' (February 2020) Cato Institute. See https://www.cato.org/blog/taiwan-election-dealing-disinformation-while-protecting-speech.

[194] Ibid.

[195] Ibid.

In terms of strict legal regulation, there are some laws that impose fines and prison time for anyone who spreads rumours.[196] According to the *Taiwan Security Brief: Disinformation, Cybersecurity, & Energy Challenges*, both the executive and legislative branches of Taiwan's government have introduced amendments to existing laws in order to limit the spread of disinformation.[197] For example, the Disaster Prevention and Protection Act was amended in to impose penalties on those who spread false information about disasters.[198]

United Kingdom

In the UK, there is no legislation that addresses disinformation, but there are a number of government initiatives. For example, the National Security Communications Team (NSCT), set up by the government has the purpose of tackling 'communications' aspects of threats to national security, such as disinformation.[199] Furthermore, the NSCT delivered a campaign called 'Don't Feed the Beast', the purpose of which was to inform the public on how they can detect disinformation before it goes viral.

The government has also announced that the intelligence services are now responsible for identifying social media platforms that distribute misinformation and disinformation under the 'Fusion Doctrine' which provides the 'Government must use the full suite of security, economic, diplomatic and influence capabilities to deliver our national security goals'. This means strategic communications are to be considered with the same seriousness as financial or military options.[200]

Moreover, the 'Rapid Response Unit' was established by Cabinet Office to help ensure that public debates are based on fact. It is made up of 'specialists including analyst-editors, data scientists, media and digital experts' who coordinate with government media teams to ensure they are equipped to quickly respond to the current news environment. The role of the Rapid Response Unit is to 'monitor news and information being shared and engaged with online to identify emerging issues with speed, accuracy and with integrity.' The Rapid Response Unit works very closely with the NSCT to provide 'highly visible public information'.[201]

In February 2019, The House of Commons Digital, Culture, Media and Sport Committee published its final report on disinformation and fake news in which it was recommended that 'clear legal liabilities should be established for technology companies to act against harmful or illegal content on their sites'.[202] It also recommended a mandatory code of ethics that is overseen by an independent regulator.[203] The UK government then stated in the *Online Harms White Paper* that it

---

[196] Lauren Dickey, 'Confronting the Challenge of Online Disinformation in Taiwan' in *Taiwan Security Brief: Disinformation, Cybersecurity, & Energy Challenges* eds Yuki Tatsumi, Pamela Kennedy, and Jason Li (September, 2019) See https://www.stimson.org/wp-content/files/file-attachments/StimsonTaiwanSecurityBrief2019.pdf.

[197] Ibid.

[198] Ibid.

[199] Clare Feikert-Ahalt, 'Government Responses to Disinformation on Social Media Platforms: United Kingdom' *Library of Congress* (September, 2009). See https://www.loc.gov/law/help/social-media-disinformation/uk.php.

[200] Ibid.

[201] Ibid.

[202] House of Commons Digital, Culture, Media and Sport Committee, *Disinformation and 'Fake News': Final Report*, Culture, Media and Sport Committee, chapter 2 [37] (February, 2019). See https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf.

[203] Ibid.

will establish a new statutory duty of care on technology companies to keep their users safe, and that this will be overseen by an independent regulator.[204]

New Zealand

The New Zealand Government does not appear to have a statutory or self-regulatory regime to combat the spread of disinformation. On 31 October 2019, it was announced that the Minister of Justice, Andrew Little had plans in place to combat this issue at this year's election.[205] This included a 'special team within the Ministry of Justice to direct people to information aimed to be as accurate and neutral as possible, and to be on the look-out for any attempts to deliberately mislead the public.' It was also stated that the Electoral Commission would look after the essential elements of running the referendums. The Electoral Commission would keep an eye to ensure that disinformation would not spread. However, the justice team would be in charge of the public information, websites, and will respond to public queries. Furthermore, the team would also have a monitoring role.[206]

In January 2020, New Zealand Prime Minister Jacinda Ardern announced that the Labour Party will voluntarily sign up to Facebook's new advertising transparency rules.[207] Facebook had introduced authorisation and transparency measures so that election voters could see who is behind paid advertising online, how much they are spending and who they are targeting.[208] These measures are mandatory in some other countries like the US, UK, Canada but New Zealand is voluntarily adopting them. The Labour party also guaranteed that the costings of all major new policy announcements released during the election campaign will be independently verified. The Government stated that it will continue to work on establishing an independent policy costing unit for the 2023 election.[209]

Canada

In May 2019, the Canadian Government launched Canada's Digital Charter.[210] This charter contains 10 points addressing a range of digital issues including privacy and data concerns, competition law concerns and approaches to digital disinformation. While this charter articulates a broad range of principles for digital safety, the final three points address digital hate speech and disinformation.[211] There is no definition of disinformation or misinformation in the Charter.[212]

---

[204] Department for Digital, Culture, Media and Sport and Home Office, *Online Harms White Paper (*April, 2019). See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

[205] Jane Patterson, 'Plans to combat misinformation in election-year referendum debates' (October, 2019) *RNZ*. See https://www.rnz.co.nz/news/political/402155/plans-to-combat-misinformation-in-election-year-referendum-debates.

[206] Ibid.

[207] Ben McKay, 'Fake news in Ardern's Sights for NZ Poll' (January, 2020) *Illawarra Mercury*. See https://www.illawarramercury.com.au/story/6594298/fake-news-in-arderns-sights-for-nz-poll/?cs=7479.

[208] Ibid.

[209] Ibid.

[210] Corinne Reichert, 'Canada launches Digital Charter to combat hate speech and fake news' (May 2019) *CNet*. See https://www.cnet.com/news/canada-launches-digital-charter-to-combat-hate-speech-and-fake-news/.

[211] Ibid.

[212] Funke and Flamini 'A Guide to anti-misinformation actions around the world'.

In January 2019, the Canadian Government announced a multi-prong effort to combat misinformation and disinformation. For instance, in relation to misinformation, it was announced that the Canadian Government would provide $7 million in funding for projects aimed at increasing public awareness of misinformation.[213] The measures combatting digital disinformation were done in line with reforms to Canada's electoral law that criminalised foreign funding of partisan advertising activity.[214] This ad registry is effectively a gallery feature which archives all paid political or partisan content hosted during the election period.[215] These measures required that digital platforms compile a political ad registry[216] and placed strict rules and spending caps on third parties involved in partisan activity. In the lead up to the 2019 elections, Google and Twitter chose not to prohibit electoral and issue advertising. Google's ban was in place until the election period concluded on 21 October, while Twitter was not accepting political and issue-based ads until the vote was called and the gallery feature was ready.[217] These programs are overseen through a new Government cyber security department.[218]

There are currently no Canadian laws that prohibit the dissemination of incorrect information.[219] Nevertheless, the measures introduced in January 2019 build on the *Elections Modernisation Act* (2018), which requires further transparency from technology companies. These requirements recognise the role that digital platforms play in modern democracies, however they aim to encourage platforms to enforce policies to limit the potential that they are manipulated to spread disinformation.[220] Specifically, technology companies are required to be more transparent in their anti-disinformation and advertising policies regarding elections.[221] So far, Twitter, Facebook and Google have committed to the gallery feature discussed above.[222]

Czech Republic

In April 2016, the Czech Interior Ministry announced the launch of the 'Centre Against Terrorism and Hybrid Threats'.[223] This Centre became operational on the 1st January 2017 and is a twenty-

---

[213] Ibid.

[214] Jillian Linton, 'Disinformation and elections: lessons from Canada' (December, 2019) *The RSA*. See https://www.thersa.org/discover/publications-and-articles/rsa-blogs/2019/12/disinformation-elections-canada.

[215] Ali Salam, 'So what are these new Elections Canada advertising rules anyway?' (August, 2019) The National. See https://www.national.ca/en/perspectives/detail/so-what-are-these-new-elections-canada-advertising-rules-anyways/.

[216] This is similar to the political ad library that Facebook began operating in the EU for the European Union elections. This registry can be accessed here: https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=AU&impression_search_field=has_impressions_lifetime.

[217] Above n 215.

[218] Jillian Linton, 'Disinformation and elections: lessons from Canada' (December, 2019) *The RSA*. See https://www.thersa.org/discover/publications-and-articles/rsa-blogs/2019/12/disinformation-elections-canada.

[219] Tariq Ahmad, 'Initiatives to counter fake news: Canada' (April 2019) *Library of Congress*. See https://www.loc.gov/law/help/fake-news/canada.php>.

[220] Government of Canada, 'Expecting social media platforms to act' (March, 2019) . See https://www.canada.ca/en/democratic-institutions/news/2019/01/encouraging-social-media-platforms-to-act.html.

[221] Funke and Flamini, 'A Guide to anti-misinformation actions around the world'.

[222] Salam, 'So what are these new Elections Canada advertising rules anyway?'

[223] Kremlin Watch 'Policy Shift Overview: How the Czech Republic became one of the European leaders in countering Russian Disinformation' (May,2017) *European Values Think-Tank*. See https://www.kremlinwatch.eu/userfiles/policy-

person team with an analytical and communications role.[224] The Centre aims to analyse trends in potential disinformation and communicate this to both the general and professional public.[225] The Centre utilises digital platforms in order to spread awareness of disinformation issues occurring within the Czech Republic. For example, the centre has a Twitter feed, which it regularly updates to flag and debunk disinformation issues.[226] As the Czech Government has classed disinformation as a potential threat to internal security, which falls within the jurisdiction of the Ministry of the Interior, this Centre has been included in this Ministry.[227] This highlights the focus of the Czech Government on disinformation as an issue aligned with foreign influence, primarily that of Russia and Russian linked organisations. These responses were enacted in the lead up to 2017 elections occurring in the Czech Republic.[228]

United States of America

The US has taken actions aimed at combatting state-based disinformation occurring in the international sphere. For example, the Global Engagement Centre, an organisation within the Department of State, was established with the aim of combatting state-sponsored disinformation.[229] This agency is similar to the Disinformation Review, established by the European Union.[230] US State Governments have also introduced programs or other initiatives in over 24 states to improve media literacy.[231] For example, in September 2018, the Californian State Government introduced measures to bolster media literacy by requiring the Department of Education to list instructional materials and resources for evaluating the trustworthiness of media online.

Under Section 230 of the *Communications Decency Act (CDA)* owners of interactive computer services are limited from liability for content generated and posted by third parties, and have protection over the removal of content in order to uphold their terms of service under what is known as a 'good Samaritan clause'. This provision was recently brought back into focus after President Trump signed an executive order declaring that platforms would need to demonstrate the 'good faith' element of their content moderation after his personal tweets attracted intervention from content moderators on Twitter.[232] Tweets posted by the President attracted a relatively new fact-checking function, which flagged the tweets and provided links to fact-checking materials. Twitter

shift-overview-how-the-czech-republic-became-one-of-the-european-leaders-in-countering-russian-disinformation_15273205250003.pdf.

[224] Ibid 12; Emily Schultheis 'Czech Republic's Fake News Problem' (October,2017) *The Atlantic*. See https://www.theatlantic.com/international/archive/2017/10/fake-news-in-the-czech-republic/543591/.

[225] Kremlin Watch 'Policy Shift Overview: How the Czech Republic became one of the European leaders in countering Russian Disinformation', p 12.

[226] Ibid.

[227] Ibid.

[228] Robert Tait, 'Czech Republic to fight "fake news" with specialist unit' (December 2016) *The Guardian* accessible at https://www.theguardian.com/media/2016/dec/28/czech-republic-to-fight-fake-news-with-specialist-unit.

[229] Alberto Alemanno 'How to counter fake news: a taxonomy of anti-fake news approaches' (2018) 9(1) *European Journal of Risk Regulation* 1, p 3.

[230] Ibid.

[231] Funke and Flamini, 'A Guide to anti-misinformation actions around the world'.

[232] Renee Diresta, 'Social Media Fact-Checking is not Censorship' (June 2020) *Slate.* See https://slate.com/technology/2020/06/twitter-fact-checking-trump-misinformation-censorship.html.

also obscured one of President Trump's tweets for violating rules around glorifying violence.[233]

There are also various bills proposed in the US to regulate certain specific aspects of misinformation. For example, a Bill for the *Deepfake Report Act* of 2019 would require the Secretary of Homeland Security via the Under Secretary for Science and Technology (S&T) to publish an annual report for the next five years on the use of deep fake or 'digital content forgery' technology. Furthermore, the *DEEP FAKES Accountability Act* (2019) is aimed at combatting the spread of disinformation through restrictions on deep fake video alteration technology. The *Honest Ads Act* was announced in 2017,[234] which is a bill that seeks to increase transparency by aligning online political advertising disclosure laws with those for radio and television. The bill would prevent foreign nationals and entities from purchasing political advertisements online, and improves transparency by expanding disclosure rules from just ads that explicitly endorse or oppose a candidate to include ads that mention a candidate, and by requiring platforms to maintain a database of online political advertisements.[235] Although this legislation was re-introduced in May 2019, it has not been passed and some commentators have noted that it 'faces long odds of becoming law'.[236]

South Korea

The South Korean Government has existing measures for blocking known disinformation originating from North Korea or North Korean aligned sources, and the dissemination of pro-North Korean propaganda is criminalised.[237] However, there have been limited moves to implement measures to prevent the spread of disinformation originating in the domestic media ecosystem.

South Korean responses have focused on 'fake news', translated literally into Korean as '*Ga-jja-new-su*' and used as a popular term for 'false or fabricated information, regardless of motive'.[238] Bills put before Parliament do not share a common definition of fake news, an issue highlighted by the Parliament's Science, ICT, Broadcasting and Communications Committee.[239] South Korea has two existing media regulatory bodies, the Korean Communications Committee and the Korean Communications Standards Committee, the latter of which oversees online and social media.[240] These two bodies can compel media organisations to issue apologies and corrections for media content that is deemed to be false. There have been some concerns raised that they can act as a form of government censorship as they are government organisations that have suppressed

[233] Anthony Zurcher, 'Trump signs executive order targeting Twitter after fact-checking row' (May 2020) *BBC.* See https://www.bbc.com/news/technology-52843986.

[234] Colin Lecher, 'Senators announce new bill that would regulate online political ads' (October 2017) *The Verge*. See https://www.theverge.com/2017/10/19/16502946/facebook-twitter-russia-honest-ads-act.

[235] Tim Lau, 'The Honest Ads Act Explained' (January 2020) The *Brennan Center.* See https://www.brennancenter.org/our-work/research-reports/honest-ads-act-explained.

[236] Zach Montellaro, 'The Honest Ads Act Returns' (May 2019) *Politico*. See https://www.politico.com/newsletters/morning-score/2019/05/09/the-honest-ads-act-returns-615586; Rachel Baker, 'Covid Communications: How Fake News Fanned Coronavirus Hysteria' (June 2020) 39(2) *Communications Law Bulletin* 1, p 3.

[237] Casey Corcoran, Bo Julie Crowley, and Raina Davis, 'Disinformation Threat Watch' (May, 2019) Harvard *Kennedy School: Belfer Center for Science and International Affairs*). See https://www.belfercenter.org/sites/default/files/2019-06/PAE/DisinfoWatch%20-%202.pdf.

[238] Kanchan Kaur et al, 'Information Disorder in Asia and the Pacific: Overview of Misinformation Ecosystem in Australia, India, Indonesia, Japan, the Philippines, Singapore, South Korea, Taiwan, and Vietnam' (March, 2019) *SSRN Online.* See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3134581, p 39.

[239] Ibid p 42.

[240] Ibid p 32.

negative stories of sitting Presidents.[241]

There are already some efforts to self-regulate fake news on digital platforms in South Korea. Online portals, including the two most popular, provided by Naver and Daum, require news providers to go through an evaluation process before they can display content.[242] The Seoul National University (SNU) runs a fact-checking initiative partnered with the news section on Naver, Korea's largest internet portal.[243] While this benefits from SNU's reputation as an independent arbiter or truth, this initiative has struggled to gain traction with low levels of internet traffic.[244]

Singapore

Singapore recently implemented legislation that gives the Government a large level of power in policing disinformation and misinformation online. In May 2019, Singapore passed the *Protection from Online Falsehoods and Manipulation Act* which criminalises the dissemination of false information online.[245] This Act gives any government Minister the power to give directions regarding information that is deemed to be a false or misleading statement of fact.[246] This can include that access to the content is disabled or a correction notice is affixed.[247] The law imposes significant penalties for breaching these provisions where a malicious actor that shares false information can face a fine of up to $37,000 or five years in prison.[248] This doubles to $74,000 or 10 years in prison if the sharing is done through an inauthentic online account or through a bot.[249] This Act has been used at least two dozen times since its first use in November 2019. This Act includes provisions for digital platforms that do not comply with Ministerial directions. Failure to comply with an order to disable access to a site can result in a fine of up to $14,400 per day for platforms.[250]

This Act has drawn criticism from human rights groups, political groups and technology companies.[251] Importantly, it has been criticised for excessive restrictions on freedom of expression that can potentially be used to stifle criticism of the Government.[252] In response, the Government

---

[241] Ibid p 32.

[242] Ibid p 43.

[243] Corcoran et al, 'Disinformation threat watch. The disinformation landscape in East Asia and implications for US policy, Technical report'. (2019). See https://www.belfercenter.org/.pdf. p 31.

[244] Ibid.

[245] Rachel Baker, 'Covid Communications: How Fake News Fanned Coronavirus Hysteria' (June 2020) 39(2) *Communications Law Bulletin* 1, 2-3; Shibani Mahtani, 'Singapore introduced tough new laws against fake news. Coronavirus has put them to the test' (March 2020) *The Washington Post*. See https://www.washingtonpost.com/world/asia_pacific/exploiting-fake-news-laws-singapore-targets-tech-firms-over-coronavirus-falsehoods/2020/03/16/a49d6aa0-5f8f-11ea-ac50-18701e14e06d_story.html.

[246] Ibid

[247] Above n 245.

[248] Funke and Flamini, 'A Guide to anti-misinformation actions around the world'.

[249] Ibid.

[250] Above n 245

[251] Ibid, Baker, 'Covid Communications: How Fake News Fanned Coronavirus Hysteria' 2-3; Tessa Wong, 'Singapore fake news law polices chats and online platforms' (May, 2020) *BBC* accessible at https://www.bbc.com/news/world-asia-48196985.

[252] Bhavan Jaipragas and Dewey Sim, 'Singapore's fake news law: protecting the truth, or restricting free debate?' (December 2019) *South China Morning Post* accessible at https://www.scmp.com/week-asia/politics/article/3043034/singapores-fake-news-law-protecting-truth-or-restricting-free.

has claimed the Act does not amount to censorship, as posts remain online with a corrections label affixed.[253] The Act has also been criticised for its broad phrasing which lacks a clear definition of false statement of fact, nor a definition of 'public interest' that is to be referred to by Ministers in their decision making.[254]

[253] Ibid.

[254] Above n 251, Wong.

Appendix

A framework for information disorder

In 'Fake News. It's Complicated'[255] First Draft founder Claire Wardle outlined seven types of mis- and dis-information (Figure 1).

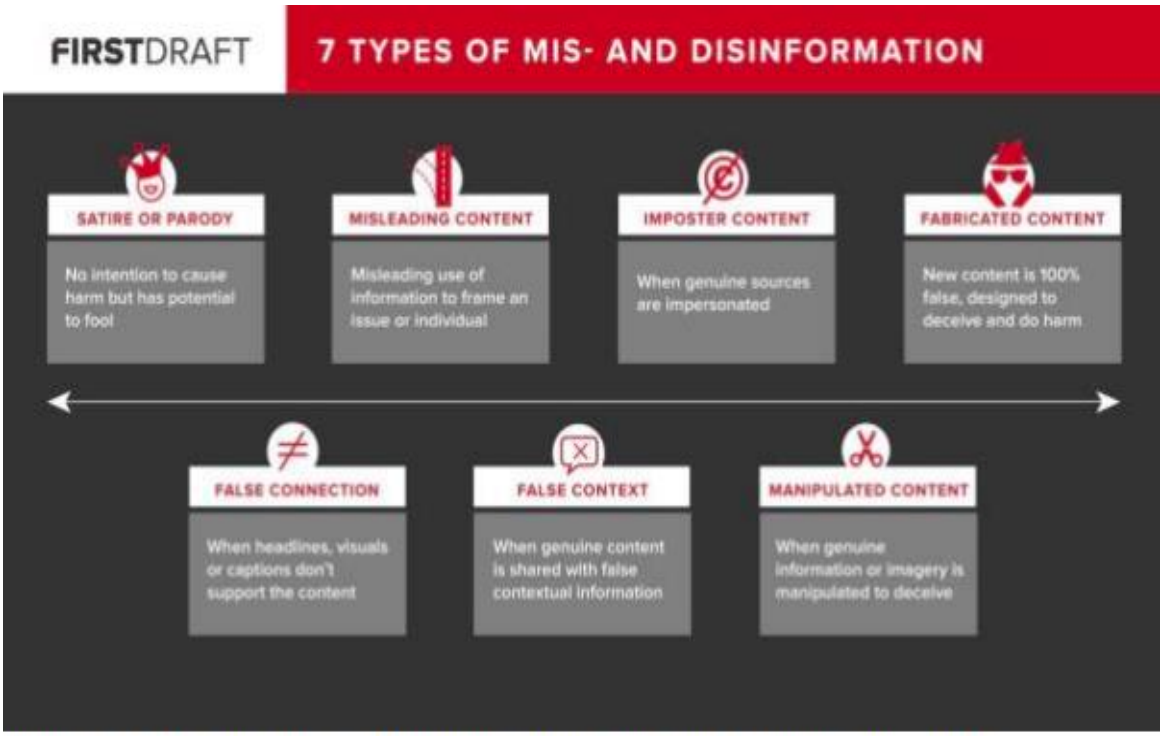


Figure 1: 7 Types of Mis- and Dis-information (Credit: Claire Wardle, First Draft)

Claire Wardle and Hossein Derakhshan further created a more nuanced framework as an organising structure of information disorder for the 2017 Council of Europe Report (COE) 'Information Disorder: Towards an Interdisciplinary Framework'.[256] This output is now widely adopted by both scholarly and industry practitioners as well a guidebook by UNESCO.[257] Wardle and Derakhshan's conceptual framework in the COE report[258] outlines three components, each of which is also broken down into three parts including the types, phases and elements of information disorder as outlined below.

[255] See https://firstdraftnews.org/latest/fake-news-complicated/.

[256] See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

[257] See https://en.unesco.org/fightfakenews.

[258] See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

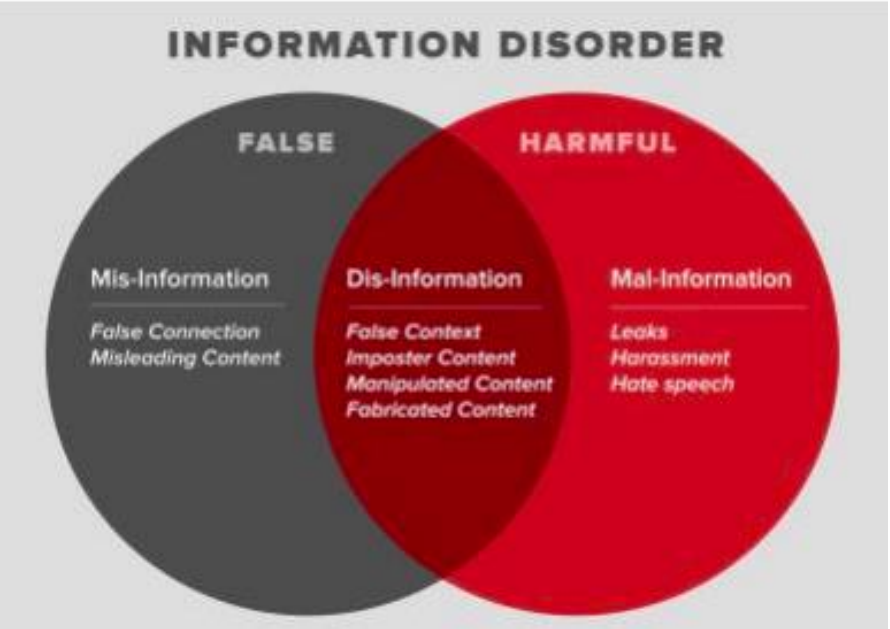


Figure 2: The Three Types of Information Disorder: Dis-information, Mis-information and Mal-information (Credit Claire Wardle and Hossein Derakhshan)

The three phases of information disorder (Figure 3) consider the cycles of creation of content and re-creation and re distribution of the content. Original messages may target an audience, but the reproduction and re distribution may be accessed by new audiences other than those originally intended.

Figure 3: Three Phases of Information Disorder (Credit Claire Wardle and Hossein Derakhshan)

A holistic view of information disorder is provided by 'The Three Elements of Information Disorder: Agent, Message and Interpreter' (Figure 4 below).

Messages can be re-shared and produced differently to how the original agent intended. This framework highlights the importance of internet users in its focus on the 'interpreter' and suggests the importance of initiatives in areas such as media literacy where the 'interpreter' learns not to 'share' misinformation and disinformation. Training in digital and media literacy initiatives to help the public at large to recognise signals that point to problematic messages can be delivered via easy

access methods[259] and warrant further research and experimentation to allow digital citizens to consider their own roles and responsibilities in curbing mis-and disinformation.

Figure 4: Three Elements of Information Disorder (Credit Claire Wardle and Hossein Derakhshan)

[259] See https://www.niemanlab.org/2020/07/first-draft-launches-a-text-message-course-to-help-inoculate-users-against-us-election-misinformation/.