7

Responses Aimed at The Target Audiences of Disinformation Campaigns

Chapter 7 of the report:

Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

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

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

ISBN 978-92-3-100403-2

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

7.1 Normative and ethical responses

Author: Julie Posetti

This chapter will discuss ethical and normative responses to disinformation executed at international, regional and local levels. These efforts frequently involve public condemnation of acts of disinformation, or recommendations and/or resolutions concerning responses. They extend to initiatives designed to embed values and actions at the individual level that can help counter the spread of disinformation. Because much disinformation may not be illegal (unless it is used for financial fraud, or incitement to violence), there is a wide realm of ethical decision-making by various actors concerning the production, hosting and sharing of fabricated information.

The triangle of norms, ethics and laws can be unpacked in various ways. In this chapter, it is understood that these elements may be aligned, or in tension with each other. Norms and ethics in some cases may run counter to legal frameworks, while personal ethics can involve individuals challenging a particular norm.

7.1.1 What are the aims of ethical and normative responses?

Ethical and normative responses to disinformation primarily operate at the level of shaping norms, thereby reinforcing a shared social assumption that disinformation is a serious threat to society. They are designed to influence individual ethical decisions to identify, counter and deter the production and distribution of disinformation.

This set of responses is not about 'external' protection of the targets or recipients of disinformation, but rather about increasing efforts to prepare people to be active agents in building their own resistance to disinformation. It assumes that the behaviours of those targeted are influenced by norms and ethics, and that the interventions will strengthen these in the interests of helping to 'inoculate' against, and collectively counter, disinformation.

The related expectation is that people are moral, rational and open to 'vaccinating' themselves against viral disinformation. Some evidence suggests, however, that many people choose to believe, endorse and circulate erroneous information that reinforces their beliefs or prejudices, in preference to engaging with accurate, credible content that may challenge them to shift their opinions and add nuance to their identities.

As discussed in chapter 3 (Research Context & Gaps) research in the fields of psychology and sociology has emphasised the complex role and functions of human cognition, belief, and social mores in the disinformation ecosystem. Falsehoods are smuggled into people's consciousness by focusing on beliefs rather than reason, and feelings instead of deduction. The spread of disinformation relies on prejudices, polarisation, partisanship, and identity politics, as well as credulity, cynicism and individuals' search for simple sense-making in the face of great complexity and change (Posetti & Bontcheva, 2020a). This explains why much research indicates, misconceptions can be extremely hard to shift, especially when identifiable facts are shrouded in untruths, even (or, perhaps, especially) when fact-checkers debunk false information. Further, as several sources

have demonstrated, repetition and rhetoric strengthen belief in inaccurate information (e.g. Zacharia, 2019). Other research has concluded that ethical concerns about sharing falsehoods are reduced with repeated exposure (Effron & Raj, 2019). Ethical and normative responses to disinformation should therefore be mindful of these complexities and structured to adapt to them.

The word 'trust' appears dozens of times in this report because many efforts to respond to disinformation are linked to the issue of trust - trust in facts, trust in reputable institutions, and trust in information sources. Trust is implicated as both a vector for disinformation and a potential antidote to it - from the problem of so called 'trust networks' (those networks of family and friends on social media) that propel disinformation (Buchanan & Benson, 2019), to disinformation-laced attempts to undermine trust in critical independent journalism, and encourage cynicism (as distinct from scepticism) and conspiracy thinking about news and truth (Ireton & Posetti, 2018; Vaccari & Chadwick, 2020). Trust is a critical but elusive ingredient in dealing with disinformation. Normative and ethical responses to disinformation impact on the issue of trust by creating a beacon or moral social compass for societal conduct in producing, transmitting, consuming and regulating content.

7.1.2 Who and what are the targets of ethical and normative responses?

The responses in this category are typically aimed at the norms and ethics of targets and recipients of disinformation. Member States of intergovernmental organisations, policy makers, and legal and judicial actors are a primary focus of these interventions. But the broad citizenry, online communities, internet communications companies, news publishers and journalists are also targeted.

These interventions rely on the extent to which those targeted are aligned to international norms on human rights (especially freedom of expression), and are also both able and willing to adhere to codes of ethics, and interested in improving their regulations, policies and practices in response to disinformation challenges.

For example, journalist-oriented initiatives operate on the assumption that journalists have the latitude and the conscience to adhere to codes of ethics (Storm, 2020) and that they are interested in improving the factual accuracy of their coverage in the face of disinformation challenges (Taylor, 2020). They also depend to an extent on which of these standards and norms are embedded within the professional context, and institutionally within news organisations.

Institutional arrangements such as self-regulatory councils are key for underpinning norms and ethics both regarding the media and the internet communications companies. One recent attempt to apply more robust self-regulatory frameworks in this realm is the Facebook Oversight Board (Clegg, 2020; Wong, 2020a). It is a formally appointed semi-autonomous board that will review decisions to remove content (notably, this will not involve informing decisions about what content is kept online in face of complaints). There is no explicit mention of the role of disinformation, misinformation or fact-checking in the Board's charter (Facebook, 2019e), nor what Facebook calls 'coordinated inauthentic behaviour' (i.e. organised misinformation and disinformation), although these may be reasons for content removal. It is not evident what norms and standards will be applied to such determinations if the Board is expected to review such decisions. On the other hand, the newly-appointed Board's Deputy Chair has publicly expressed a desire to "audit" Facdbook fact-checking efforts as part of the Oversight Board's work, stating that

that there are "serious concerns" about political bias in fact-checking and questioning the commitment of fact-checkers to the "facts" (Allen, 2020).

The norms shaping governmental engagement with disinformation are similarly linked to institutional underpinnings such as parliaments, courts and independent communications regulators.

7.1.3 Who are the primary actors and what responses do they produce?

The main actors initiating normative and ethical responses to disinformation are: intergovernmental organisations at the international level (e.g. UNESCO, WHO, UNDP) and regional levels (e.g. EU, CoE, OAS, AU); internet communications companies; news organisations; journalists; and civil society organisations. Below, specific examples of these responses are catalogued and analysed.

a. Intergovernmental responses

At the intergovernmental organisation level, there have been several noteworthy recommendations, statements, and reports produced in an effort to reinforce values and frameworks designed to counter disinformation within the boundaries of international human rights law.

In a significant development in June 2020, a cross-regional statement was issued by more than 130 UN member states and official observers, in the context of COVID-19. This statement said: "It is critical States counter misinformation as a toxic driver of secondary impacts of the pandemic that can heighten the risk of conflict, violence, human rights violations and mass atrocities. For these reasons we call on everybody to immediately cease spreading misinformation… ." The statement further noted "…the key role of free, independent, responsible and pluralistic media to enhance transparency, accountability and trust, which is essential to achieving adequate support for, and compliance by, the general public with collective efforts to curb the spread of the virus". In calling on countries to take steps to counter the spread of such disinformation, the statement advised that efforts should be based on "freedom of expression, freedom of the press and promotion of highest ethics and standards of the press, the protection of journalists and other media workers, as well as promoting information and media literacy, public trust in science, facts, independent media, state and international institutions" (UN Africa Renewal, 2020).

United Nations level responses

Another UN normative intervention is the 2017 'Joint Declaration On Freedom Of Expression and Fake News", Disinformation and Propaganda' (OSCE, 2017). This declaration was issued by the United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information.

This joint statement, produced in collaboration with the civil society organisations Article 19 and the Centre for Law and Democracy, came in response to a rash of legislation from multiple states seeking to address the disinformation crisis by prohibiting the

publication and dissemination of certain content. It seeks to address both the causes and consequences of disinformation (including both disinformation-fuelled attacks on the news media by state actors, and the rush to regulate against disinformation) through the framework of international human rights law, emphasising enshrined freedom of expression rights. The statement indicates that the signatories are:

> *"Alarmed at instances in which public authorities denigrate, intimidate and threaten the media, including by stating that the media is "the opposition" or is "lying" and has a hidden political agenda, which increases the risk of threats and violence against journalists, undermines public trust and confidence in journalism as a public watchdog, and may mislead the public by blurring the lines between disinformation and media products containing independently verifiable facts."*

Recognising the potential for so called 'fake news legislation' to infringe on freedom of expression rights, in particular through inadvertently (or by design) curtailment and suppression of legitimate journalism, it also emphasises that:

> *"...the human right to impart information and ideas is not limited to "correct" statements, that the right also protects information and ideas that may shock, offend and disturb, and that prohibitions on disinformation may violate international human rights standards, while, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or State actors."*

The objective of such statements is to sensitise UN Member States about their responsibilities under international human rights law, and to encourage adherence as a way of dissuading both the use of disinformation as a tool to intimidate or regulate the news media and other publishers of public interest information as a means of limiting freedom of expression. The target audiences of such interventions also include policy makers, the news media, and by extension, the broader public.

Associated approaches to reinforcing normative values and ethical standards adopted by UN agencies include UNESCO's #MILCLICKS campaign and its 'Journalism, 'Fake News' and Disinformation' handbook (Ireton & Posetti, 2018). The former initiative seeks to foster Media and Information Literacy[356] (MIL), through Critical-thinking, Creativity, Literacy, Intercultural, Citizenship, Knowledge and Sustainability (CLICKS).[357] It is aimed at young audiences, and is designed to foster critical engagement with information online - a cornerstone of medium and longer term responses to disinformation - promoting the notion of #ThinkBeforeSharing. The normative practice being encouraged is accountability for communications, as well as informed and ethical reflection about how individuals engage with content. The UNESCO handbook, meanwhile, is aimed at embedding ethical, accountable and critical approaches to combatting disinformation within journalism education and training. The handbook adopts an ethical framework for journalism's defence against disinformation: "Ethical journalism that values transparent practice and accountability is a vital piece of the armoury in the battle to defend facts and truth in an era of 'information disorder'." (Ireton & Posetti, 2018). It further elaborates:

356 https://en.unesco.org/themes/media-and-information-literacy
357 https://en.unesco.org/MILCLICKS

> *Professional standards for ethical and accountable journalism are an important defence against disinformation and misinformation. Norms and values providing guidance to people doing journalism have evolved over the years to give journalism its distinctive mission and modus operandi. In turn, these uphold verifiable information and informed comment shared in the public interest. It is these factors that underpin the credibility of journalism. As such, they are woven into the fabric of this handbook. (Ireton & Posetti, 2018)*

Regional level responses

Policy initiatives, charters of obligations, inquiries and targeted research from the European Commission and the Council of Europe have contributed to a comprehensive attempt to reinforce normative and ethical responses to disinformation in Europe.

The European Commission promotes the normative understanding that disinformation can "cause public harm, be a threat to democratic political and policy-making processes, and may even put the protection of EU citizens' health, security and their environment at risk." (European Commission, 2019). It outlines its policy approach and intent regarding efforts to combat disinformation in its online policy repository, with objectives summarised thus:

> *The exposure of citizens to large scale disinformation, including misleading or outright false information, is a major challenge for Europe. The Commission is working to implement a clear, comprehensive and broad set of actions to tackle the spread and impact of online disinformation in Europe and ensure the protection of European values and democratic systems. (European Commission, 2019).*

This approach has been informed by collaborative scholarship and expert consultations, including the work of the EU's High Level Expert Group on 'Fake News' and Online Disinformation. In its final report (Buning et al., 2018), the Group made a series of recommendations that emphasise the values of privacy, professional ethics, and social responsibility.

One initiative to flow from the European Commission's normative policy approach (European Commission, 2018a) is an Action Plan Against Disinformation (European Commission, 2018e) which is designed to deal with legal acts of disinformation and is couched in terms of geopolitical threats and the need to reinforce European democratic values: "This Action Plan was a response in 2019 to the European Council's call for measures to 'protect the Union's democratic systems and combat disinformation, including in the context of the upcoming European elections'."

Another action-oriented outcome focused on ethics is the European Commission's Code of Practice on Disinformation (European Commission, 2018c), which was published in late 2018 with the assertion that: "This is the first time worldwide that industry agrees, on a voluntary basis, to self-regulatory standards to fight disinformation." Signatories now include Facebook, Google, Twitter, Mozilla and Microsoft, along with eight advertising trade associations (European Commission, 2018d). Stated objectives of the Code include transparency in political advertising, although there is no reference to accuracy or fact-checking associated with political advertising. This is relevant to ongoing debates

connected to political advertising during elections, which have seen calls for the introduction of 'truth in political advertising' standards.[358]

The Council of Europe commissioned a foundational research report which emphasises the role of professional ethics and norms in combatting what it defines as 'information disorder' (Wardle & Derakhshan, 2017). The report offers a range of recommendations for states, technology companies, the news media, civil society and funders. Beyond disruption to democratic elections, the report identified the biggest concern demanding attention as: "...the long-term implications of disinformation campaigns designed specifically to sow mistrust and confusion and to sharpen existing sociocultural divisions using nationalistic, ethnic, racial and religious tensions." (p. 4) This points to the need for responses to disinformation that recognise the risks at the intersection with hate speech and seek to reinforce norms and values like racial and gender equality, and religious tolerance.

The 2017 "Joint Declaration On Freedom Of Expression and 'Fake News', Disinformation and Propaganda", referenced above, was also signed by regional intergovernmental organisations representing Latin America and Africa, along with OSCE. The OSCE Representative on Freedom of the Media has also reiterated: "...at all times, and especially in difficult times, blocking or banning media outlets is not an answer to the phenomenon of disinformation and propaganda, as it leads to arbitrary and politically motivated actions. Limits on media freedom for the sake of political expediency lead to censorship and, when begun, censorship never stops. Instead, the answer lies in more debate and media pluralism".[359] Additionally, the OSCE has supported country-specific workshops designed to embed freedom of expression norms in responses to disinformation while practically equipping Member States to respond to disinformation (OSCE, 2017).

b. Civil Society responses

Many civil society responses to disinformation involve initiatives that seek to reinforce democratic values and human rights frameworks that support norms like freedom of expression, access to information, privacy and gender and racial equality. Several of these interventions, operating at the intersection of disinformation and hate speech, are detailed in section d. below.

Many of the examples of Media and Information Literacy initiatives from civil society organisations identified in the next chapter are also designed with strong normative and ethical components at the core. Such initiatives seek to stimulate grassroots ethical responses to disinformation.

One notable civil society initiative designed to address information pollution is Reporters Without Borders' (RSF) Forum on Information and Democracy[360] based upon an international declaration endorsed by 38 countries. This initiative evaluates norms and architectures of global communications networks, investigates companies' actions, makes recommendations, facilitates regulation and self-regulation, commissions research and supports journalism.

358 See discussion below

359 https://www.osce.org/fom/319286

360 https://informationdemocracy.org/; disclosure: the author of this chapter is a member of the steering committee of the Forum's Working Group on Infodemics https://informationdemocracy.org/working-groups/concrete-solutions-against-the-infodemic/

c. Responses from the internet communications and news industries

From the internet communications companies to news organisations, a range of normative and ethical responses to disinformation can be catalogued.

Twitter decided to ban political candidate advertising from its site ahead of the 2019 UK elections, with company CEO and founder Jack Dorsey announcing via a tweet: "We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought." (Dorsey, 2019). Google followed suit a couple of weeks later, replicating Twitter's commitment to prevent micro-targeting of users for politically-themed adverts. Additionally, Google promised to ban 'deepfakes' and what it termed "demonstrably false claims" to try to protect the integrity of elections and support trust in democratic processes (Wong, 2019a).

As a result, Facebook came under mounting ethical pressure to address its policies pertaining to misinformation and disinformation connected to political advertising and speech on its site (see chapter 4.1) - brought into sharp focus by the Cambridge Analytica scandal[361] - after it decided not to apply fact-checking standards to certain types of political advertising (Eisenstat, 2019; Stewart, 2019). Facebook considered restricting the micro-targeting of users by political actors (Glazer, 2019). However, the company ultimately announced that it would not curtail such micro-targeting, and that no action would be taken to prevent politicians from making false claims in their posts, nor in paid advertising, ahead of the 2020 U.S. election (Romm et al., 2020). Under this policy (see chapter 4.1), the company further excluded certain types of political advertising content from the fact-checking work which it contracts out (meaning therefore that it also does not label this kind of content as false and misleading) (Hern 2019a; Van Den Berg & Snelderwaard, 2019) . However, the company did proceed with new protocols in the U.S. that meant it could ask its fact-checking partners to assess the truthfulness of non-political advertising on Facebook (Hern, 2019b).

Facebook's normative argument is that, in general, it is inappropriate for a private company to be an arbiter of truth in the case of political advertising (Gilbert, 2019). In a 2019 blog post, Facebook's Vice-President for Global Affairs and Communications, Nick Clegg, argued that freedom of expression is "an absolute founding principle for Facebook" (Clegg, 2019). As noted by UN Rapporteur on Freedom of Opinion and Expression, David Kaye, avoiding being an arbiter of truth should not exclude Facebook from taking any action against clear falsehoods (Kaye, 2020b). The normative debate in practice is balancing the company's interpretation of freedom of expression with actual limitations on expression set out in the company's community standards, and how these limits compare to those permissible for states to make under international human rights law. The result is controversy over whether cases violate Facebook's own community standards or raise issues of restriction under international standards (which the private sector is expected to respect, according to the UN's principles agreed in the Ruggie Report[362]). An example is conspiracy theories, which in principle are tolerated on the service, unless these are deemed to contain false or misleading content that can cause imminent harm. There was, however, evidence of a more restrictive approach emerging in mid 2020, when Facebook removed nearly 800 pages and groups, and restricted approximately 2000 Instagram accounts in connection with the QAnon conspiracy theory. (Facebook, 2020b)

361 https://www.theguardian.com/news/series/cambridge-analytica-files
362 https://www.business-humanrights.org

Ethical concerns about Facebook's approach to fact-checking political advertising motivated hundreds of the company's employees to argue in a letter to management that: "Free speech and paid speech are not the same thing." They claimed that policies on avoiding fact checking advertisements from politicians, political parties and their affiliates "are a threat to what FB stands for". They stated that the policy does not protect voices, but instead "allows politicians to weaponize our platform by targeting people who believe that content posted by political figures is trustworthy." (New York Times, 2019)

It is important to note, however, that Facebook policy still allows the rejection of direct speech or advertising by political candidates, incumbents, political parties and their affiliates if it amounts to an immediate threat to safety "in the real world", or if it contravenes the company's voter suppression policies (Facebook, 2019d). For example, on March 30th, 2020, Facebook and Instagram removed videos of Brazilian president Jair Bolsonaro for spreading disinformation on the coronavirus and therefore violating the platforms' terms of use. Those terms do not allow "misinformation" that could cause physical harm to individuals, said Facebook (BBC News, 2020b). However, these standards are not applied uniformly internationally. For example, posts quoting U.S. President Donald Trump speculating on bleach as a potential treatment for COVID-19 were not removed (Suárez, 2020).

Although Facebook CEO Mark Zuckerberg was cited stating that promoting bleach as a cure for coronavirus was the kind of "misinformation" that would be removed immediately - because of "imminent risk of danger" - the company said that Trump's statement did not violate the policy because he did not specifically direct people to ingest bleach. Since then, Facebook has removed a video in which the U.S. President claimed children were "virtually immune" to coronavirus (Kang and Frenkel 2020). The issue of Facebook applying its standards differently around the world has been recognised by former senior Facebook policy manager Richard Allan, who explained differences in treatment in terms of "risk" related to the proximity of a country to the U.S. and its size. (Suárez, 2020). In September 2020, Buzzfeed published extracts from a memo by a former Facebook data scientist who claimed that outside of Western countries, the company regularly abrogated its responsibility to deal with political disinformation with the potential to cause physical harm. She cited instances in multiple developing countries. (Silverman, Mac and Dixit, 2020).

Related concerns were also raised in a UK House of Lords report which assessed that "Facebook have purposefully hobbled their third-party fact checking initiative by exempting all elected politicians and candidates for office from being fact checked." (House of Lords, 2020).

Deciding when content is opinion or fact when these are closely intertwined in a given item requires, inter alia, an ethical judgement call. As discussed in chapter 4.1, this highlights policy loopholes whereby disinformation may not be labelled as such, or fact-checking labels denoting falsity are removed by the company, because falsehoods are bundled with opinion (which Facebook policy regards as largely exempt from fact-checking), thereby creating conundrums for what constitutes an appropriate response at an ethical level. For example, Facebook has removed fact-checking labels applied by third party fact-checkers to content deemed to be opinion (Penney, 2020; Pasternack, 2020; Grossman & Schickler, 2019). There are also reports of pressure being applied by the company to third party fact-checkers in reference to the fact-checkers' assessment of opinion and 'advocacy' content, and fact-checkers being wrongly accused of bias with regard to labelling scientific disinformation, with very limited transparency (Pasternak, 2020).

While Facebook has long-running formal fact-checking partnerships with many reputable news organisations and NGOs,[363] several of which have described a mission-driven motivation for participating (Funke & Mantzarlis, 2018a), the initiative has attracted ethical critiques from some journalists. Those actively engaged in third party fact-checking who feel that the collaboration clashed with professional norms have been among these critics (Levin, 2018). A number of fact-checking partners have ultimately pulled out of the arrangement in the midst of debates on professional ethics connected to the operation of Facebook's Third Party Fact-Checking Program (Lee, 2019a). Among them was U.S.-based anti-hoax website Snopes. One of Snopes.com founders indicated that the ethical challenges were among the reasons for withdrawing (Green, 2019). Facebook's fact-checking partner in the Netherlands, Nu.nl, also withdrew from the project[364]. The non-profit outlet took a values-based decision to quit the collaboration in disagreement with Facebook's adoption of an ethical position to exempt political advertisements (with some exceptions) from its fact-checking (Hern, 2019a; Van Den Berg & Snelderwaard, 2019).

Another example of a news organisation demonstrating competing norms is the BBC's complaint about a Facebook advertisement which used a decontextualised and misleading clip of Political Editor Laura Kuenssberg appearing to endorse the Conservative Party's Brexit strategy. This presented significant reputational and ethical challenges for a public broadcaster that holds up political neutrality as one of its core values. Facebook banned the advertisement several days after receiving the complaint, with the justification that this was a copyright breach (Mays, 2019). At the time it was banned, GBP 5000 had been spent on the advertising campaign which had appeared in news feeds about 250,000 times (Who Targets Me, 2019).

More recently, Facebook moved to thwart politically affiliated publishers masquerading as local news sites from claiming exemption from the company's political advertisement authorisation process (Fisher, 2020). This followed publication of research from the Tow Center for Digital Journalism that revealed over 1200 cases of political groups posing as local news sites to publish propaganda in the U.S.. (Bengani, 2020).

Additional ethical and normative responses to disinformation have come from some news organisations putting disinformation combat at the core of their editorial strategies. For example, a 2019 study (Posetti et al., 2019a) identified a 'mission-driven' approach to combating disinformation from three Global South news organisations: Rappler in The Philippines, the Daily Maverick in South Africa, and The Quint in India. Each of these news organisations identified a commitment to reinforcing democratic principles, defending media freedom, and adhering to the core ethical tenet of 'speaking truth to power' in response to state-sponsored disinformation networks and foreign influence agents that they believed were destabilising their democracies. Additionally, they sought to model these norms for their audiences as a means of motivating the ethical responsibility to eschew disinformation practices, including attacks against journalists laced with 'lies'. One example of this approach is audience-focused campaigns from Rappler, encouraging the community to join them in opposing online hate connected to orchestrated disinformation campaigns which targeted Rappler and its CEO-Executive Editor, Maria Ressa (Posetti et al., 2019b). These were operationalised online using hashtags like #NoPlaceForHate, #IStandWithRappler and #HoldTheLine, the objective being to demonstrate a shared ethical commitment to combating disinformation within online communities and opposing state-based disinformation campaigns as being antithetical to cultural and social norms and mores.

363 See detailed analysis in Chapter 4

364 Facebook's Third Party fact-checking programme has since relaunched in the country with two partners: AFP and DPA

In 2020, in calling on Facebook to assume moral responsibility to act in response to disinformation, Ressa cited the UN's conclusion that Facebook had played a "determining role" [365] in what the UN has described as a "a textbook example of ethnic cleansing"[366] against the Rohingya in Myanmar through its facilitation of both disinformation and hate speech (Posetti, 2020). Facebook later acknowledged that "we weren't doing enough to help prevent our platform from being used to foment division and incite offline violence," and said it had updated its policies to "now remove misinformation that has the potential to contribute to imminent violence or physical harm."[367]

The South African National Editors' Forum (SANEF) has also played a normative role in highlighting the dangers of disinformation and working closely with Media Monitoring Africa to confirm a commitment to the eradication of disinformation[368]. Initiatives connected to this collaboration include the disinformation reporting portal called Real411.

Another collaborative response to disinformation from the journalism community came during the 2019 World Media Summit (an initiative of China's Xinhua news agency, which now involves 13 international media partners, including Reuters, BBC, and AP). The Summit reportedly reached a consensus on disinformation: "To ensure the authority and credibility of media are upheld, media have the mission to fight against disinformation; false information should be clarified without delay; and fake news should be boycotted by all... . The reporting and spreading of fake news violate journalistic ethics and damage the interests of the general public" (Xinhua, 2019).

d. Anti-hate speech initiatives

Where disinformation intersects with hate speech - such as racism, misogyny and bigotry - normative and ethical responses are often triggered. These span initiatives from civil society organisations, and intergovernmental agencies through to interventions from celebrities. One such celebrity is comedian-actor Sacha Baron Cohen, whose speech on social media-fuelled disinformation and propaganda to an Anti-Defamation League conference on antisemitism and hate in November 2019 sought to get the companies involved to take action against disinformation endangering religious and ethnic minorities (Baron Cohen, 2019).

There are also interventions from research institutes and NGOs seeking to provide normative guidance through development of frameworks designed to embed values-based approaches to managing hate speech as it manifests as a feature of orchestrated disinformation campaigns. One example of such an intervention is the International Foundation for Electoral Systems' exploration of the links between hate speech and disinformation, and provision of a normative framework for programming interventions (Reppell & Shein, 2019).

Another example is an RSF report 'Attack of the Trolls' that covers the online abuse of journalists - particularly at the intersection of misogyny and disinformation. It sought to raise awareness and activate responses designed to reinforce press freedom norms online (RSF, 2018) Similarly, the FOJO Media Institute's #JournoDefender initiative[369] focused on combatting online misogyny as it intersects with disinformation fuelled-attacks designed

365 https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf
366 https://news.un.org/en/story/2017/09/564622-un-human-rights-chief-points-textbook-example-ethnic-cleansing-myanmar
367 https://about.fb.com/news/2018/11/myanmar-hria/
368 https://sanef.org.za/disinformation/
369 https://journodefender.org/

to undermine democracy. This initiative was underpinned by research conducted in multiple countries (FOJO: Media Institute, 2018).

Journalists and news organisations have themselves sought to reinforce values of gender equality by investigating disinformation campaigns involving misogynistic elements. For example, Rappler Editor and CEO Maria Ressa cites a commitment to the principles of 'speaking truth to power' and 'shining a light' as reasons she chose to speak out publicly about her experience of being brutally harassed online in retaliation for investigative journalism that exposed reportedly government-linked disinformation networks in the Philippines. (GIJN Staff, 2019)

Finally, UN Special Rapporteurs have signalled online hate-speech deploying disinformation tactics against female journalists. Five UN Special Rapporteurs issued a joint statement in 2018 calling on the Indian Government to protect Indian journalist Rana Ayuub who was bombarded with death threats as part of a misogynistic disinformation campaign which used 'deepfake' videos and fake accounts to misrepresent her and expose her to risk (UN Human Rights, 2018).

7.1.4 Response Case Study: COVID-19 Disinformation

Ethical and normative responses include public condemnation of acts of disinformation, or recommendations and resolutions aimed at thwarting these acts because of the life-threatening character of the pandemic. Such responses include statements from UN special rapporteurs, WHO officials, and national leaders. Additionally, there have been examples of calls for reinforcing ethical conduct within journalism, and for internet communications companies to do more. These responses have often taken the form of published statements, speeches or articles designed to move others to stop sharing disinformation, to reinforce freedom of expression norms during the crisis, and to adapt ethical standards to address new challenges in responses to what two UNESCO-commissioned policy briefs responding to the COVID-19 disinformation crisis framed as the 'disinfodemic' (Posetti & Bontcheva, 2020a; Posetti & Bontcheva, 2020b).

Examples include:

- A World Press Freedom Day statement from UN Secretary General Antonio Gutteres reinforcing the normative role of professional journalism in the information ecosystem as a bulwark against disinformation. This statement also asserted the ethical and legal obligations of UN Member States regarding press freedom rights (and journalism safety mechanisms) in the context of responses to COVID-19. (UN Secretary General, 2020)

- A joint statement from International experts including David Kaye, UN Special Rapporteur on the right to Freedom of Opinion and Expression; Harlem Désir, OSCE Representative on Freedom of the Media, and Edison Lanza, IACHR Special Rapporteur for Freedom of Expression: "Governments must promote and protect access to and free flow of information during pandemic". (UN Human Rights, 2020a)

- A report to the UN Human Rights Council from UN Special Rapporteur on the right to Freedom of Opinion and Expression David Kaye which explicitly appealed to the moral and ethical obligations and responsibilities of Member States in reference to their COVID-19 responses (Kaye, 2020a). The report states that it is a "plea to all

Governments to treat those within their jurisdictions ... with the dignity and respect demanded by international human rights law."

- Calls from senior editors, journalists and media academics to stop live broadcasting politicians who disseminate disinformation during speeches and press conferences, due to the difficulty of fact-checking and debunking in real-time. (Thomas, 2020)

- Unprecedented decisions by internet communications companies to edit or remove recordings of political leaders deemed to be spreading disinformation about COVID-19. (BBC, 2020b)

- As noted above, the crisis triggered more than 130 United Nations member countries and official observers to urge that all steps to counter COVID-19 should be based, inter alia, on respect or freedom of expression and press freedom.[370]

- The African Commission on Human and Peoples' Rights issued a press statement on a "human rights based effective response" to the pandemic. This reiterated the obligation of States to ensure that the measures adopted comply with the principle of legality, are necessary and proportional to the objective of safeguarding public health. Such measures include those intended "to dispel misinformation and myths about COVID19 and to penalize the dissemination of false information on risks of COVID19".[371]

7.1.5 How are ethical and normative responses evaluated?

There is very limited evidence of any kind of evaluation associated with ethical and normative responses to disinformation, in part because of the methodological difficulty of such an exercise. One contributing factor is that embedding ethics and norms within societies, or stimulating commitments to international human rights principles is a highly collaborative process and it is close to impossible to determine which actor, or which particular message, was more or less transformative.

The relevant UN Special Rapporteur monitors Member States' handling of disinformation in reference to their adherence to international human rights norms like freedom of expression, and issues assessments to the UN Human Rights Council on that basis (UN Human Rights, 2020b). But there is no known evaluative process that seeks to directly attribute the development of norms and ethics within societies to such interventions. Some case references however do show impact in certain contexts.[372]

For example, a statement from the UN Secretary General (UN Secretary General, 2020) highlighting the fundamental importance of ensuring that counter-disinformation measures introduced by Member States do not undermine the principles like press freedom, is typically amplified by the news media (Apelblat, 2020) and reinforced by civil society organisations' efforts to embed norms like 'access to information' (Article 19, 2020). However, beyond media measurement exercises by commercial service providers,

370 https://www.un.org/africarenewal/news/coronavirus/cross-regional-statement-%E2%80%9Cinfodemic%E2%80%9D-context-covid-19
371 https://www.achpr.org/pressrelease/detail?id=483
372 https://www.lefigaro.fr/flash-actu/la-bolivie-abroge-des-decrets-anti-desinformation-controverses-20200515

there is not systematic and at scale publicly-available research about the extent of dissemination and amplification of these kinds of statements.

7.1.6 Challenges and opportunities

These normative and ethical interventions can be comparatively simple and affordable to implement and they can work as counter-narratives that appeal to individuals' moral compasses, or reinforce alignment with values like anti-racism or anti-misogyny. A problem, however, is when moral compasses and societal norms are not linked to the principles of access to information, freedom of expression, press freedom, and privacy - as enshrined in international human rights law. There are many attempts to normalise the expression and dissemination of false and misleading content that is potentially harmful.

One of the most significant risks associated with state-based responses to disinformation is posed by legal and regulatory approaches that go against the international norms of freedom of expression (including its corollary press freedom) and privacy.

As highlighted by the rush of responses to the disinfodemic that accompanied the COVID-19 pandemic, a disinformation crisis can lead to changes in what is accepted as normal, such as the suspension or weakening of human rights protections (Posetti & Bontcheva 2020a; Posetti & Bontcheva 2020b). Further, even though restrictions on the right to seek, receive and impart content can be legitimate under international standards, when these are for reasons of public health, it remains the case that they need to be in law, as well as necessary and proportional to the purpose.

In such circumstances, responses like 'fake news' new laws that effectively criminalise journalism can exceed these standards, and also go on to become entrenched as new norms. It is therefore a challenge to ensure that all interventions are anchored within the legal and normative frameworks of human rights, and particularly freedom of expression (including press freedom and access to information) and privacy.

Ultimately, legitimate normative and ethical responses to disinformation can be de-legitimised by the individuals, organisations and States who disagree with the intention behind them, in the same way that credible journalism can be misrepresented as 'fake news' by those seeking to avoid being held to account.

Many actors highlighting these issues seek to address the challenge of a risk of downgrading human rights standards by empowering the public (and their elected representatives) to recognise that such interventions against disinformation (where they are justified during emergencies like the COVID-19 pandemic) should inter alia be of time-limited duration. However, the impact of these messages depends on persuading those with power to tack closely to these standards.

The main opportunity in the category of ethical and normative responses to disinformation is to reaffirm and remind people about norms around access to information and freedom of expression. In the COVID-19 crisis, it can be underlined that these norms are not only about fundamental rights, but also significant tools for mitigating impact and tackling disinformation.

Immediate normative steps to counter disinformation can also be taken with an eye to promoting long-term normative and institutional impacts in terms of international standards. For example, news reporting on disinformation responses can explain the

importance of assessing these initiatives against international human rights laws, and the normative and ethical frameworks that support them. Such explanatory journalism could aid accountability on the part of governments and corporate actors, as well as help embed understanding of the role of these values and standards in areas beyond dealing with disinformation.

7.1.7 Recommendations for normative and ethical responses

The challenges and opportunities identified above, and their significant implications for freedom of expression, provide a basis for options for action in this category of responses.

Recommendations for action in this chapter include strengthening the institutional underpinnings for freedom of expression norms, as components of disinformation responses. In this regard:

International organizations could:

- Conduct follow-up evaluation of the circulation of, and engagement with, normative statements as well as assessment of the actual impact of ethical codes, such as operated by internet communications companies and news media that are relevant to disinformation issues.

Individual states could:

- Ensure institutionalised multi-stakeholder governance of internet communications companies, covering transparency and the range of policies on disinformation in the context of content curation.
- Embed human rights impacts assessments within responses to disinformation from executive or legislative branches of government, especially those which risk overreach (e.g. the expansion of 'fake news' laws in the context of the COVID-19 pandemic).

Internet communications companies could:

- Commit to values that defend vulnerable communities and groups, including from threats in multiple languages, and ensure that all countries in which they operate are served by measures adopted to combat disinformation.
- Engage diverse stakeholders in developing policies that support ethical decision-making concerning disinformation content - including if it should be removed.
- Increase capacity to deal with disinformation at scale, especially in countries in conflict, and provide swift responses to actors targeted by this disinformation, as well as redress opportunities in regard to decisions on how relevant content is treated.

- Strengthen their normative role regarding freedom of expression by ensuring regular independent review of their disinformation-related policies and implementation, and the human rights impacts thereof.

- Recognise that an ethical commitment to freedom of expression does not preclude a range of decisive actions on political disinformation that is likely to cause significant harm - such as where it threatens lives, public health, the institutions of democracy, or electoral integrity.

- Enhance transparency and disclosure of data about practical processes around managing disinformation.

Media actors could:

- Ensure that they adhere to the highest ethical and professional standards to avoid becoming captured or associated with disinformation purveyors.

- Invest in investigative journalism focused on exposing disinformation networks and explaining the risks of disinformation to their audiences and the importance of resisting it in the public interest, as a means of building trust while also pursuing truth.

- Increase the capacity of independent press councils to monitor and address disinformation (including when it spreads through news media channels) and disinformation responses (especially as they affect freedom of expression) as part of their ethics oversight role.

Researchers could:

- Use audience research methods to measure the influence and impact of messaging aimed at developing ethics and values that help inoculate against disinformation, or undertake qualitative research into normative evolution and behavioural change focused on disinformation defences.

- Study Media and Information Literacy initiatives to assess the impact on participants' behaviours and sense of personal accountability regarding the need to counter disinformation.

7.2 Educational responses

Authors: Kalina Bontcheva, Julie Posetti and Denis Teyssou

Educational responses are aimed at improving citizens' media and information literacy and promoting critical thinking and verification in the context of online information consumption, as well as journalism training and tools designed to strengthen fact-checking, verification, and debunking.

Of particular relevance are critical thinking, news and advertising literacy, human rights awareness, identity issues, understanding of algorithms and data, and knowledge of the political economy of communications (including economics, psychology and sociology around the production, targeting and spread of disinformation).

This section provides an overview of different kinds of educational responses, distinguished by cataloguing the organisations that design and deliver them and identifying the targets of these responses. In conclusion, they are assessed as to how they address disinformation in relation to educating learners about the fundamental value of freedom of expression, and explain the difference between mobilising and interpreting different facts on the one hand (which would not constitute disinformation), and on the other, when false or misleading information is mobilised and interpreted (which is the essence of disinformation).

7.2.1 What and who do educational responses target?

Media and information literacy and critical thinking initiatives are widely regarded as key 21st century skills, required by citizens to more effectively discern and counter online disinformation. As noted in a report by the Broadband Commission: "Traditional school curricula tend to prioritize the accumulation of knowledge over the application of knowledge, and many school systems fail to adequately train students in digital citizenship and literacy." (Broadband Commission, 2013).

The notion of Media and Information Literacy (MIL) as UNESCO uses it, includes a range of competencies concerning the consumption, production and circulation of content. Under this umbrella are knowledge and skills covering fields such as critical thinking, content production; news literacy; advertising literacy; film literacy; political economy of communications; algorithmic literacy; privacy literacy; and intercultural communication (Berger, 2019). UNESCO also operates with a concept of Global Citizenship Education[373] (GCED), which includes competencies around identity and values. Together, these represent a "playbook" that can help empower participants in digital communications to deal with disinformation in a range of ways. Educational initiatives in the wide field of MIL may be formal and informal, and spread across a range of social institutions from schools through to cities, transportation systems, as well as media and social media.

UNESCO's wide range of target competencies emphasises the comprehensive breadth required for MIL efforts to be successful. While many efforts tend to focus on news and

[373] https://en.unesco.org/themes/gced

verification literacies alone, the strongly interconnected topics of algorithmic, advertising, and privacy literacy are very rarely addressed. The notion of "digital literacy " is variably elaborated as to what competencies it aims to prescribe. As argued by some civil society organisations[374], it also is very important to educate children (and adults) about how personal data are collected and shared online for commercial gain; the hidden dangers of online profiling and targeting; algorithms and their biases; and user privacy online. Thus an important gap in a number of MIL toolkits and programmes is in the lack of coverage on the concept of data literacy. Meanwhile, data literacy in the face of disinformation links closely to the issue of digital citizenship (Carmi et al., 2020).

Addressing these MIL challenges through long-lasting, effective educational responses is a key part of the puzzle, since research has found in some instances that the main amplifiers behind viral disinformation are human users (Vosoughi et al., 2018). The key questions, then, are why do citizens 'fall for' online disinformation, what motivates them to share it (even if they are aware it is untrue), and what is the impact of online disinformation on their offline behaviour (e.g. does it affect their voting in elections)? Particularly in the context of COVID-19, many citizens are being duped and are propagating online disinformation, leaving them unable to understand and implement scientifically-grounded preventive measures. People are dying as a result of complacency (Karimi & Gambrell, 2020), or resorting to false 'cures' (Embury-Dennis, 2020).

Both scientists (e.g. Corbu et al., 2020) and fact-checkers (Vicol, 2020) have been studying the question of what makes citizens believe and spread false or misleading content . Age, education, and repetitive exposure to disinformation have all been confirmed as important factors (Vicol, 2020), with adults over 50 and those without higher education being particularly at risk. Another important factor is *confirmation bias*, i.e. people's tendency to read and believe content which conforms to their existing worldviews (Nickerson, 1998; Corbu et al.,2020; Nygren & Guath, 2019). According to a study by Gallup and the Knight Foundation (Knight Foundation, 2018), people generally share information that they trust and do so primarily for social or personal reasons. Moreover, an individual's online news and information sharing and commenting behaviour is influenced by the behaviour of their typically like-minded online social connections - referred to as *homophily* in scientific studies (Tarbush & Teytelboym, 2012). Receiving content from trusted sources such as friends and families adds credence to the credibility of this content.

Possibly linked to all this, researchers have found that "social networks and search engines are associated with an increase in the mean ideological distance between individuals" (Flaxman et al., 2018), i.e. lead to polarisation. These findings hold across many countries (Kelly & François, 2018).

Experimental research has also shown that when polarised online communities are exposed to disinformation which conforms to their preferred narratives, it is believed and shared (Quattrociocchi et al., 2016). Consequently, when such users and communities are exposed to debunks or opposing opinions, these may either have little effect or can even strengthen their pre-existing beliefs and misconceptions. Moreover, a recent FullFact survey showed that homophily motivated 25% of UK adults to share content even though they believed it to be made up or exaggerated (Vicol, 2020).

Researchers from the YouCheck! MIL project have also found evidence of overconfidence (Nygren & Guath, 2019; Nygren & Guath, 2020) and a `third person effect' (Durach,

[374] https://5rightsfoundation.com/our-work/data-literacy/

2020; Corbu et al., 2020), where people rate their own capabilities to detect online disinformation too favourably compared to the abilities of others.

Taken together, this evidence demonstrates the key importance of developing effective MIL and GECD responses to disinformation.

7.2.2 Who do educational responses try to help?

MIL and GCED are widely regarded as key skills that enable citizens to discern online disinformation more effectively. Where citizen-oriented surveys have been carried out, however, evidence has emerged consistently that the majority of citizens are lacking these essential skills. For instance, a 2018 Eurobarometer survey (Eurobarometer, 2018) in the 28 EU member states established that only 15% of the respondents felt very confident in identifying online disinformation. Other surveys focused specifically on the citizen's ability to distinguish factual from opinion statements, where a Pew Research Center study (Mitchell et al., 2018) has shown that (on average) only 26% of Americans were able to recognise factual news statements, with the number rising to 33% for younger Americans. In addition, a RISJ report (Newman, 2019b) findings have indicated a global tendency to conflate poor journalism with disinformation and 'fake news'.

This has motivated the emergence of initiatives aimed at improving media, digital and data literacy, and critical thinking across all ages (from school children, through to retirees). Data literacy in the face of disinformation links closely to the issue of digital citizenship (Carmi et al, 2020).

Complementing these efforts are initiatives and resources, aimed at educating journalism students and professional journalists in the most up-to-date tools, methodologies, and resources for verifying, investigating and debunking online disinformation. These are often developed and facilitated by leading journalists, journalism educators, researchers, and civil society organisations. Frequently, these efforts are also highly collaborative.

7.2.3 What output do educational responses publish?

Outputs Aimed at Improving Citizen's MIL and GCED

One class of media literacy initiatives relies on learning through **online games**, i.e. teaching citizens media literacy and critical thinking through participation in a game. This is an engaging way for people (not just school children) to gain knowledge and experience. One example is the Drog initiative[375] which has brought together academics, journalists, and media experts to build an online game - GetBadNews. The game aims to educate people about the various tactics employed in online propaganda and disinformation campaigns. Another educational game is Fakey[376] by the University of Indiana, which asks players to share or like credible articles and report for fact-checking suspicious ones. The BBC has developed the iReporter[377] interactive game (Scott, 2018), which gives young players the role of a journalist who needs to report on news without falling prey to disinformation. Another notable example is the multilingual YouCheck!

375 https://aboutbadnews.com/
376 https://fakey.iuni.iu.edu/
377 https://www.bbc.co.uk/news/resources/idt-8760dd58-84f9-4c98-ade2-590562670096

Detectives[378] fake news game, which is available in English, French, Spanish, Romanian, and Swedish. For older teens (15-18 years old), the global International Factchecking Network has produced a role-playing card game (currently in English, Italian, Portuguese, and Spanish),[379] with students playing newsroom journalists covering a controversial referendum, marred by online propaganda and disinformation. Other examples have been collected by the American Press Institute (Funke & Benkelman, 2019). The connection between MIL and games has also been recognised by UNESCO and selected governments in a pioneering Games Bar session held in late 2019.[380]

There are also more traditional, **school-based approaches** to media literacy, which are targeting pre-teens and teens, just as they start taking interest in social media, news, and politics. A prominent government-led response comes from Finland, where the public education system has media literacy classes as standard (Henley, 2020). This is reported to have made Finnish citizens well prepared to recognise online falsehoods. Elsewhere, media organisations and civil society groups are filling the gap in state-led provision in schools. Examples include the school media club run by the NGO African Centre for Media and Information Literacy (AFRICMIL)[381] and the MENA student-oriented MIL activities of the Media and Digital Literacy Academy of Beirut (MDLAB).[382] Another example is Lie Detectors[383] - a non-profit initiative in Belgium and Germany, which puts journalists in the classrooms to interact directly with pupils and teach about news literacy and news verification practices. There is a similar initiative in France led by journalists from Le Monde (Roucaute, 2017) and the Redes Cordiais journalist-led initiative in Brazil.[384]

A more global initiative comes from the BBC, who with the help of the British Council, are providing global media literacy resources for schools around the world (BBC, 2020a). Elsewhere, politicians have spearheaded a programme to teach media literacy to high school children, an initiative that can have a bearing on building resilience to disinformation (Troop, 2017). Another example is Poynter's MediaWise initiative[385], which has delivered MIL face-to-face and online training to over five million teenagers and other citizens, with special focus on under-served communities.

A complementary activity to school-based MIL approaches is **teacher training**, i.e. MIL training aimed at training teachers who can then in turn deliver successful school-based MIL training to students. UNESCO has several resources here and a global process to revise and update a curriculum framework for teachers in the light of recent developments such as the proliferation of disinformation.[386] This is an essential as it enables school-based MIL training to scale up and become sustainable. Examples include the Brazilian Educamidia project[387] and the European YouCheck! project.[388] In France, work with school teachers on the problem of "infox" takes place with the Savoir*Devenir initiative among others.[389]

378 http://project-youcheck.com/game-french/
379 https://factcheckingday.com/lesson-plan
380 https://en.unesco.org/news/media-and-information-literacy-joins-games-learning
381 https://www.africmil.org/programmes-and-projects/media-information-literacy/school-media-clubmedia-in-education/
382 https://mdlab.lau.edu.lb
383 https://lie-detectors.org/
384 https://www.redescordiais.com.br/
385 https://www.poynter.org/mediawise/
386 https://en.unesco.org/sites/default/files/belgrade_recommendations_on_draft_global_standards_for_mil_curricula_guidelines_12_november.pdf
387 https://educamidia.org.br/habilidades
388 http://project-youcheck.com/about/
389 http://savoirdevenir.net/mediatropismes/

Media-content approaches: recent studies show that the older generation (above 50 years old) in some countries has lower than average ability to recognise factual information (Gottfried & Grieco, 2018) and to remember already debunked false claims (Mantzarlis, 2017). One significant challenge is therefore how best to deliver media and information competencies to that target demographic. The previously discussed approaches are not suitable, as older people are much less likely to use online games and rely significantly less on social media platforms as their source of news (Ofcom, 2018a). One promising way is to deliver special programmes through mainstream TV channels. As part of the Beyond Fake News project, the BBC has developed an entire series of documentaries, special reports and features across the BBC's networks in Africa, India, Asia Pacific, Europe, and the Americas which are delivered via TV, radio, and online (BBC, 2018c).

MIL training for online influencers and youth organisations: Digital influencers with their millions of followers have the propensity to spread disinformation widely and thus journalists in Brazil have started dedicated MIL training initiatives aimed to improve the ability of these celebrities to fact-check online content prior to publishing posts in support of false or misleading content (Estarque, 2020). The delivery of MIL training through youth organisations is another promising approach that is being explored with the support of UNESCO in India, Kenya, and Nigeria.[390]

Online verification toolkits and educational materials aimed at improving the general public's understanding of verification and fact checking are also increasingly becoming available e.g. Edutopia[391], a New York Times lesson plan (Schulten & Brown, 2017). UNESCO's MIL CLICKS[392] campaign and its MOOCs promoting media and information literacy, critical thinking, creativity, citizenship, and related skills, with materials in multiple languages (e.g. Arabic, English, French, Greek, Spanish). With support from the UN Office on Drugs and Crime, work was carried out in South Africa on localising a UN model curriculum on disinformation and ethics[393].

In an attempt to make content verification easier to understand, the International Fact Checking Network (IFCN) has produced a 7-step fact checking cartoon[394], currently available in English, French, Italian, Japanese, Portuguese, Serbian, and Swahili. The UK independent fact-checking charity FullFact has produced a similar 10-step misinformation detection toolkit[395], as well as offering a collection of child-oriented literacy materials. Another example is an online educational video "How to spot fake news" by FactCheck.org[396].

Outputs Aimed at Improving Journalistic Professionalism

Firstly, a growing number of **journalist-oriented verification literacy materials** and programmes isbeing created, e.g. the learning module on the history of disinformation (Posetti & Matthews, 2018). UNESCO's Journalism, *Fake News and* Disinformation *Handbook for Journalism Education and Training* (Ireton & Posetti, 2018) is available as a

390 http://www.unesco.org/new/en/communication-and-information/resources/news-and-in-focus-articles/all-news/news/unesco_supported_mil_training_in_india_three_days_of_learni/
391 https://www.edutopia.org/blogs/tag/media-literacy
392 https://en.unesco.org/MILCLICKS
393 http://www.cfms.uct.ac.za/news/media-ethics-workshop-localizes-un-curriculum
394 https://factcheckingday.com/articles/24/this-cartoon-has-7-tips-for-fact-checking-online-information
395 https://fullfact.org/toolkit/
396 https://www.youtube.com/watch?reload=9&v=AkwWcHekMdo&feature=youtu.be

free resource in 11 languages with 30 more translations pending at the time of writing.[397] First Draft also provides courses for journalists in verifying media, websites, visual memes, and manipulated videos[398]. Most recently, they launched a coronavirus-specific resource page too[399]. These are complemented by the latest edition of the "Verification Handbook" (Silverman, 2020), which provides guidance on investigating manipulated content, platforms, and disinformation campaigns.

A second type of shared **resources for journalists** is aimed at **strengthening accuracy in reporting**. Examples include providing a trustworthy resource of curating the latest research on key news topics,[400] current advice on media engagement strategies,[401] a centralised resource of public government data[402] or thoroughly fact-checked information and statistics on the economy, healthcare, immigration, etc.[403] Many of these resources, however, are currently country- and language-specific and are designed for manual human consumption. Their usefulness in fact-checking and content verification can be improved further, if they are also made machine readable/accessible following established data interchange standards.

There is also now awareness that journalists can benefit from the latest academic **research** in the field of disinformation, and even begin to collaborate with researchers, in order to integrate findings from the latest advances in psychology, social science, and data science (Lazer et al., 2017). There is also scope for learning from experts in information influence and strategic communications (Jeangène Vilmer et al., 2018), e.g. around the best debunking strategies for countering disinformation.

As disinformation increases in volume and complexity, journalists increasingly also need help with **learning about newly emerging OSINT[404] (Open Source Intelligence) and content verification tools** and the best practices for their use. Some organisations have now started sharing lists of recommended tools and their usage, e.g. First Draft's tools collection[405], India's BusinessWorld[406]. Widely used specialised tools (e.g. the InVID/WeVerify verification plugin[407]) have also started producing online video tutorials and documentation, to enable journalists to learn the techniques and adopt them quickly in their work. A lack of funding has limited these materials from becoming accessible to journalists in multiple languages.

7.2.4 Response Case Study: COVID-19 Disinformation

In the context of the COVID-19 'disinfodemic', many educational measures are being delivered digitally - often using the same online environments where disinformation proliferates (e.g. social media). These responses are being rolled out especially by MIL

397 https://en.unesco.org/fightfakenews
398 https://firstdraftnews.org/en/education/learn/
399 https://firstdraftnews.org/long-form-article/coronavirus-resources-for-reporters/
400 https://journalistsresource.org/
401 https://mediaengagement.org/
402 https://datausa.io/
403 https://fullfact.org/finder/
404 https://en.wikipedia.org/wiki/Open-source_intelligence
405 https://firstdraftnews.org/long-form-article/coronavirus-tools-and-guides-for-journalists/
406 http://www.businessworld.in/article/5-Tools-Every-Young-Journalist-Should-Learn-About-To-Identify-Fake-News/01-04-2019-16868
407 The InVID/WeVerify verification plugin now offers online tutorials, a virtual classroom, and interactive help: https://weverify.eu/verification-plugin/

projects around the world, media, journalism-oriented civil society organisations and journalism schools, as well as governments.

Examples of media and information literacy projects include:

- Pakistan's *Dawn* newspaper has published a short citizens' guide to surviving the disinfodemic as an act of digital media literacy (Jahangir, 2020).
- The London School of Economics (LSE) has published a guide to helping children navigate COVID-19 disinformation for families forced by the pandemic to homeschool their children (Livingstone, 2020)

Educational interventions aimed at journalists focus on verification, fact-checking and ethical health reporting. Some examples:

- A free online course[408] training journalists how best to cover the pandemic has been developed by the Knight Center for Journalism in the Americas, in partnership with the World Health Organisation (WHO) and UNESCO, with support from the Knight Foundation and the United Nations Development Program (UNDP).
- First Draft's Coronavirus Information Resources page includes a 'debunk database', a curated list of sources, educational webinars about reporting on the pandemic, and tools and guides to aid COVID-19 verification and debunking.
- The African Centre for Media Excellence (ACME) hosts a curated list of resources, tools, tips and sources connected to reporting COVID-19, including a fact-checking collection.
- Afghan NGO NAI has produced "Essentials of journalism performances during COVID 19".[409]
- The Data and Society research group has produced a sheet of 10 tips for journalists covering disinformation.[410]

Of particular importance are cross-border initiatives, such as International Center for Journalists (Barnathan, 2020) with a Global Health Crisis Reporting Forum which includes an interactive, multilingual hub for thousands of journalists around the world. This aims to: aid informed, ethical reporting through direct access to credible sources of scientific and medical expertise; facilitate knowledge sharing and collaborative fact-checking/ debunking in reference to COVID-19.

7.2.5 Who are the primary actors behind educational responses and who funds them?

Multi-stakeholder Partnerships: These are MIL initiatives where multiple actors from different categories work together in a partnership. Examples include UNESCO's

408 https://www.ejta.eu/news/free-online-course-journalism-pandemic-covering-covid-19-now-and-future
409 https://nai.org.af/law-and-legal-documents/
410 https://datasociety.net/wp-content/uploads/2020/04/10-Tips-pdf.pdf

MIL global alliance "GAPMIL"[411] and its partnership with Twitter during the annual Media and Information Literacy Week[412], and the AI and media integrity work of the Partnership on AI[413], which comprises over 100 organisations, including all major internet communications companies, some major media organisations, research centres and non-profits. Another example is the MisinfoCon[414] global movement which is specifically concerned with creating tools for verification and fact checking. Its supporters organise tool demos, hackathons, talks, and discussions, including literacy and critical thinking topics.

Civil society organisations and grassroots initiatives: These are MIL programmes and resources created by non-profit organisations and/or citizens. In addition to the examples already discussed above (e.g. First Draft, Drog, LieDetectors), others include the UNESCO-chair supported Savoir*Devenir[415]; the 5Rights foundation with their focus on children data literacy[416]; the Mafindo[417] grassroots Indonesian anti-hoax project; the Google-funded Center for Digital Literacy (CDL)[418] training teacher and school children in Republic of Korea; involvement of youth groups in pan-European MIL projects (e.g. INEDU[419]); grassroots actors producing debunking videos and explainers[420].

Fact-checking Organisations and Networks also provide (mainly journalist-oriented) training sessions and publish training resources, either as individual organisations or through joint initiatives[421]. International fact-checking networks (e.g. the International Fact-Checking Network[422] (IFCN), the First Draft Partner Network[423]) and journalist organisations (e.g. the International Centre for Journalists[424] (ICFJ). Such initiatives frequently attract funding from internet communications companies.

Media organisations are also very active in the development and delivery of MIL, not only through traditional (e.g. TV) and social media channels (e.g. YouTube), but also through direct engagement (e.g. in classrooms or through journalism-oriented training workshops and events). Some examples were discussed in Section 7.2.3 above. Others include the journalist training and education work done by The African Network of Centres for Investigative Reporting (ANCIR[425]) and Code for Africa (CfA[426]);

Government-led initiatives: Many governments have now started running or supporting MIL efforts focused on disinformation. Examples of such initiatives include (many of them

411 https://en.unesco.org/themes/media-and-information-literacy/gapmil/about
412 https://en.unesco.org/news/unesco-partners-twitter-global-media-and-information-literacy-week-2018
413 https://www.partnershiponai.org/ai-and-media-integrity-steering-committee/
414 https://misinfocon.com/join-the-misinfocon-movement-f62172ccb1b
415 http://savoirdevenir.net/chaireunesco/objectifs-missions/
416 https://5rightsfoundation.com/our-work/data-literacy/
417 https://www.mafindo.or.id/about/
418 https://www.blog.google/outreach-initiatives/google-org/digital-and-media-literacy-education-korea/
419 https://in-eduproject.eu/
420 The series is called Smarter EveryDay, ran by the YouTuber engineer Detin Sandlin: https://www.youtube.com/watch?v=1PGm8LslEb4; https://www.youtube.com/watch?v=V-1RhQ1uuQ4; https://www.youtube.com/watch?v=FY_NtO7SIrY
421 Full Fact, Africa Check, and Chequado: https://fullfact.org/blog/2020/feb/joint-research-fight-bad-information/
422 https://www.poynter.org/ifcn/
423 https://firstdraftnews.org/about/
424 https://www.icfj.org/our-work
425 https://investigativecenters.org/
426 https://medium.com/code-for-africa

collated by Poynter[427]) Australia, Belgium, Canada, Denmark, Finland, France, India,[428] Netherlands, Nigeria, Singapore, Sweden, and the U.S.. One example is France's Centre de liaison de l'enseignement et des médias d'information (CLEMI) initiative[429], which perhaps uniquely involves libraries and librarians as key stakeholders in the MIL response. Working at a pan-European level, between 2016 and 2018 the European Union funded 10 projects on MIL and GECD, with more under negotiation from their 2019 funding call. The majority of these were aimed at citizens (e.g. YouCheck!), with the rest targeting journalists and news production (e.g. the The European Media Literacy Toolkit for Newsrooms).

Internet Communication Companies: Educational initiatives undertaken by these companies are aimed at:

- Teaching children MIL skills, e.g. Google's Be Internet Legends[430] and the related YouTube Be Internet Citizens[431] initiatives; Google's global Be Internet Awesome[432] initiative (currently with local resources for Argentina, Belgium, Brazil, Chile, Columbia, Italy, Mexico, Peru, Poland, Saudi Arabia, United Kingdom, and United States);

- Training journalists, improving their technology and skills, and investing in media literacy-oriented editorial projects e.g. the Facebook Journalism Project[433], the Google News Initiative[434] and the related YouTube initiative[435], Google's Fact Check Explorer[436].

7.2.6 How are educational responses evaluated?

Evaluating the success of MIL and GECD initiatives in changing citizen's disinformation consumption and sharing behaviour, is a challenging and largely open problem. According to evidence reviewed by research for this study, it appears that standard metrics and evaluation methodologies are still lacking in maturity. In particular, the challenge is to move beyond the awareness-raising stage, towards sustained and institutionalised MIL interventions that lead to measurable, lasting changes in citizens' online behaviour.

There is also the need for independent evaluation of the impartiality and comprehensiveness of MIL materials and training, in particular those created by the internet communications companies. Concerns have been raised by civil society organisations (5Rights Foundation, 2019) that these tend to focus on making users (especially children but also journalists) focused on false content at the expense of privacy issues, and rather than investing in efforts to fix these problems themselves. Deficiencies

427 https://www.poynter.org/ifcn/anti-misinformation-actions/
428 https://mgiep.unesco.org/
429 http://www.clemi.org/
430 https://beinternetlegends.withgoogle.com/en-gb
431 https://internetcitizens.withyoutube.com
432 https://beinternetawesome.withgoogle.com/en_us
433 https://www.facebook.com/facebookmedia/solutions/facebook-journalism-project
434 https://newsinitiative.withgoogle.com/dnifund/report/european-innovation-supporting-quality-journalism/
435 https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html
436 https://toolbox.google.com/factcheck/explorer

in comprehensiveness and transparency have also been flagged (5Rights Foundation, 2019) with respect to inadequate discussion of the risks arising from algorithmic profiling, automatic content moderation and amplification, and privacy implications of data collection. Similar concerns exist in connection with support from these commercial actors for educational responses designed to strengthen journalism and improve journalists' skills as a response to disinformation.

Further research is needed to study on a large, cross-platform scale, the matter of citizens' exposure and propagation of online disinformation, and on probing the impacts on citizens' understanding and experience of other kinds of disinformation responses. Of particular importance is gauging citizens' knowledge and understanding of the platforms' own algorithmic and curatorial responses and how these impact on disinformation, freedom of expression, right to privacy, and right to information. Due to the recent nature of large-scale online disinformation `wildfires', there is not an extensive body of research with answers to these key questions, and findings are geographically limited and, in some cases, somewhat contradictory. This has motivated policy makers and independent experts to recommend that governments need to invest in further research on these topics (HLEG report, 2018; DCMS report, 2018c; NED, 2018), including not just data science approaches, but also ethnographic studies (NED, 2018).

Researchers have also raised concerns about the recent tendency of focusing MIL primarily on critical thinking about news (Frau-Meigs, 2019). In particular, the concerns are that "they attract funds that could otherwise be attributed to full-fledged MIL projects; they provide one-shot school interventions without much follow-up; they do not scale-up to a national level and reach a limited amount of students."

7.2.7 Challenges and opportunities

An overall challenge is how to help the general public (especially those holding polarised views) to see the value of MIL and invest the time to learn and practice mindful social media engagement behaviour. In addition, MIL faces limits if it does not go wider than news, fact-checking and content verification, without holistic encompassing of wider digital citizenship skills - including freedom of expression and other online and offline freedoms (Frau-Meigs, 2019).

With respect to MIL and GECD responses targeting children, the main challenge is in designing content and delivery mechanisms which are sufficiently engaging and have a lasting impact, as by their very nature, child-oriented responses need to target medium - to long-term outcomes. There is also a challenge to situate MIL initiatives that target disinformation and promote critical thinking within the wider context of online safety education. For instance, there is a need to make the link between the dangers of believing and sharing disinformation on one hand, and the related wider dangers of profiling, micro-targeted advertising, or sharing GPS location on the other.

Similarly, there is a need for appropriate training and education on the professional risks faced by journalists who are frequent targets of disinformation campaigns. These campaigns typically deploy misogyny, racism and bigotry as online abuse tactics designed to discredit and mislead and they require holistic responses that address digital safety and security (Posetti, 2018a).

A target group especially under-served by MIL campaigns is that of older citizens, who according to some research are also more susceptible to believing and spreading

disinformation than other age groups (Vicol, 2020). At the same time, their use, knowledge, and understanding of social platforms can also be quite limited, which adds to the challenge of how best to design and deliver MIL campaigns effectively.

Another challenge that needs to be addressed through educational initiatives is to create awareness of the potentially negative impact of the use of automation in online platforms, namely that automated disinformation moderation techniques employed in some online environments can suffer from algorithmic bias and may discriminate against a specific user group (e.g. girls, racial or ethnic groups). Recent research (5Rights Foundation, 2019) has found that 83% of 11-12 year olds are in favour of platforms automatically removing content by default, without need for it to be flagged by a user. It is unclear however, what proportion of these children are also aware of the freedom of expression implications of unmoderated use of such automation.

A new challenge relates to the COVID-19 crisis. As noted earlier in this study, the World Health Organisation has signalled an "infodemic" meaning an overabundance of content that makes it hard for people to find trustworthy sources and reliable guidance. "[437] In this context of such content overload about the pandemic, the challenge is to develop the capacity of audiences to discern the difference between verified and false or misleading content, as well as to recognise content that is in the process of scientific assessment and validation and thus not yet in either category of true or false. A related challenge is that the educational reactions to the disinfodemic risk being exclusively short-term in focus, losing sight of possible links to long-term and institution-based empowerment programmes and policies to build MIL, including for children and older people, in relation to disinformation in general.

On the opportunity angle, the pandemic has also presented a new focal point for news media and journalists to demonstrate and explain their distinctive role, and a unique moment to sensitise citizens about freedom of expression rights and obligations, provide education to help them, and reinforce MIL and GECD knowledge and skills.

There is also an opportunity that immediate educational initiatives aimed at countering the disinfodemic can be taken with an eye to long-term educational impacts. They can be explicitly structured to ensure lasting MIL outcomes, not only specifically to COVID-19 but also other kinds of health and political or climate disinformation. The crisis provides possibilities for the public to learn to approach most content with scepticism, not cynicism, and to be empowered to make informed judgements about the 'disinfodemic' and the responses to it.

In conclusion, both a massive challenge and a major opportunity that needs to be addressed is that of making MIL and GECD education accessible to children worldwide, estimated to constitute one third of internet users globally and the generation that will in time take charge of informational and other issues (Livingstone et al., 2016). This would require governments around the world to make MIL an integral part of their national school curricula; to invest in professional training of their teachers in MIL; and to work closely with civil society and, media organisations, independent fact-checkers, and the internet communication companies in order to ensure a fully comprehensive, multi-stakeholder media and information literacy provision.

[437] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6

7.2.8 Recommendations for educational responses

The challenges and opportunities identified above, and their significant implications for freedom of expression, provide a basis for the following options for action in this category of responses.

International organisations could:

- Work towards provision of Media and Information Literacy (MIL) educational initiatives and materials aimed at currently under-served countries, languages, and demographics.
- Encourage an holistic approach to MIL that covers freedom of expression issues, as well as disinformation across different topics (such as health, politics and the environment).
- Encourage donors to invest specifically in countermeasures to disinformation that strengthen MIL (as well as freedom of expression, independent journalism, and media development).

Internet communications companies could:

- Integrate MIL into the use of their services, and empower users to understand the full range of issues relevant to disinformation, including fact-checking, algorithmic and labelling issues.
- Foster interdisciplinary action research projects designed to experiment with educational responses to disinformation, and report on these experiments in robust ways that aid knowledge sharing - both across academic disciplines and between industry, educators and researchers.
- Support the development of global and/or regional MIL responses, especially by funding projects in currently under-served regions.[438]

Individual states could:

- Put in place or strengthen MIL policies and resource allocation, especially in the educational system where teachers also need to be trained to deliver MIL to children and youth as a counter to disinformation.
- Earmark funding and support for interventions for older citizens who are both a key voter demographic and a primary risk group for spreading disinformation.
- Support initiatives to address disinformation that targets children, youth, women, unemployed people, refugees and migrants, and rural communities.

News media could:

- Use their platforms to proactively train audiences and internet users about the difference between verified information on the one hand and disinformation on

[438] See, for example: https://www.ictworks.org/african-digital-literacy-project-grant-funding/

the other, and help cultivate the requisite skills to recognise this and navigate the wider content ecosystem, along with the freedom of expression issues involved.

- Support advanced training in verification and counter-disinformation investigative techniques for editorial staff

- Collaborate with journalism schools on counter-disinformation projects involving both researchers and students to improve the capabilities of graduates and deepen their own understanding and practice

Civil society could:

- Increase work in MIL innovation such as anti-disinformation games, and develop creative ways to empower constituencies beyond the educational system who are at risk from disinformation.

- Support the development of global and/or regional MIL responses, especially in currently under-served regions.

- Provide independent evaluation of MIL initiatives carried out and/or supported by internet communications companies.

Researchers could:

- Develop and apply metrics for studying MIL in relation to disinformation.

- Focus on interdisciplinary research to develop new approaches to education as a counter disinformation measure, e.g. integrate methods from journalism studies, computer science, psychology, sociology etc.

- Forge partnerships with news organisations to help strengthen investigative reporting into disinformation and deepen audience insights with reference to engagement with counter-disinformation content.

7.3 Empowerment & credibility labelling responses

Authors: Diana Maynard, Denis Teyssou and Sam Gregory

Educational responses focus on teaching people about the importance of critical thinking and self-awareness in terms of information consumption, thereby giving them internal mental competencies. This chapter looks at empowerment responses that focus specifically on external methods, tools and websites to assist users in the actual understanding of the nature of information and its sources. Thus the two kinds of responses go hand in hand.

As discussed in the previous chapter, teaching media and information literacy to both journalists and citizens alike is one of the significant responses in play. Even if disinformation cannot be wholly thwarted, its dissemination and impact can be reduced if people are able to employ critical thinking in their media and information consumption. This competence underpins the ability to effectively recognise disinformation, along with its appeal and the sources that may promote it. Such awareness can enlist those exposed to falsity to understand their part in preventing its spread and influence.

This chapter complements the educational response focus by examining the efforts around content verification tools and web content indicators that can be seen as aids or prompts that work with people's competencies in the face of disinformation. These tools and cues are intended to help citizens and journalists to avoid falling prey to online disinformation, and to encourage good practices among journalists as well as internet and media companies when publishing information.

This also includes efforts by the news media to boost their credibility over less reliable sources, through highlighting reliable brands and public service broadcasting, as well as methods aimed at consumers for assessing and rating the credibility and reliability of news sources. Examples include Newsguard[439], Decodex[440], the Global Disinformation Index[441], the transparency standards of the Trust Project[442], and a number of browser extensions (many of which are discussed in other chapters of this report, and which are, however, external to consumption of content via apps such as Facebook). Also discussed is the recent emergence of "controlled capture" apps (e.g. TruePic[443]) and newly developed approaches to authentication and provenance tracking that are being considered for use by both individuals and media entities. These include work by the News Provenance Project (NPP)[444], the Content Authenticity Initiative[445], and the complementary Journalism Trust Initiative (JTI)[446].

439 https://www.newsguardtech.com/
440 https://www.lemonde.fr/verification/
441 https://disinformationindex.org/
442 https://thetrustproject.org/
443 https://truepic.com/
444 https://www.newsprovenanceproject.com/
445 https://theblog.adobe.com/content-authenticity-initiative
446 https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation

Bridging from MIL competencies to providing aids to users, are examples like Full Fact's educational toolkit.[447] This resource provides methodologies and suggestions for tools to assist with grasping where news is from, what is missing, and how it makes the reader feel. The initiatives, such as those described here, offer assistance which enables the user to implement these ideas, for example by making it practical to discover what the original source of a piece of information is, and how trustworthy it might be.

Unlike the kinds of fact-checking tools described in Chapter 6.2, which try to prevent the spread of disinformation directly, or which (try to) give a specific answer to the question "is this information true?", empowerment and credibility tools as discussed here instead typically put the onus on the consumer to interpret explicit signals that are given about the content. For example, signals provided by groups such as Credder and Newsguard (described below) provide information about the reliability or credibility of a source, and do not provide answers to whether a specific piece of information is true or not. Similarly, provenance-tracking initiatives show where, and in some cases how, a piece of information originated, but leave it to the user to interpret that (for example, by understanding that if a picture was taken several years ago, it may not be relevant to an event which has just happened, or has been reported as such). There are thus differences to the tools described in the previous section, as will be discussed in more detail in Section 7.3.3.

Evidence about whether something is accurate and credible is often linked to knowing who originally created the content or first shared it. Newsrooms, and people relying on social media for information, need to be investigating the source, almost before they look at the content itself. For example, people should be routinely researching the date and location where this is embedded in it.

As discussed throughout this report, and especially in Chapter 3, disinformation is frequently associated with both domestic and foreign political campaigns, and can lead to widespread mistrust in state authority. One way in which states can both allay the public's fear, and help them to distinguish which information is trustworthy, thereby providing an antidote to some disinformation, is through greater transparency. A strong regime of proactive disclosure by states, along with an effective real time information dispensation, together constitute a buttress to fortify clarity of provenance. However, vigilance must still be maintained because provenance does not equal facticity or comprehensivity. When states do not fully and promptly disclose for example the range of COVID-19 statistics on the channels that are recognisably their own, this is an invitation for understandable rumour and speculation, but also for inauthentic sources to fill the gap with disinformative content.

7.3.1 What and who do empowerment and credibility labelling responses monitor/target?

Provenance-tracking initiatives aim to assist news consumers in understanding the source of information, and thus to be more aware of misleading information, which is complementary to efforts that try to prevent it occurring and spreading in the first place. These initiatives treat attribution information and metadata as tools that can give insight. This is typically relevant to fake images and videos (such as deepfakes, but also ones

[447] https://fullfact.org/toolkit/

which have just been falsely attributed) by means of better authentication. Examples include those by Twitter, Truepic, Serelay, and Amber (further detail below). Alternatively, they may target the ways in which information is displayed to the user, in order to alert them to potentially misleading information such as old content that has (been) resurfaced as if it is current, and sources that might be dubious or untrustworthy. Examples of this include the News Provenance Project and Adobe, as well as Twitter's policy on dealing with manipulated media, which provides in-platform context (Roth & Achuthan, 2020).

Trust-based initiatives, on the other hand, monitor the news providers themselves, attempting to distinguish those which are unreliable, as well as sometimes assessing individual articles and authors. Accreditation-based initiatives largely target the news providers, aiming to "legitimise" those which are most trustworthy. Some, such as Newsguard, also target social media providers and other intermediaries' media, in the hope of financial returns for providing these companies with tools to rank news providers and sources.

7.3.2 Who do empowering and credibility labelling responses try to help?

These initiatives aim at five main types of audience. The majority of them try to help the general public by making them more aware of potential issues, but some also target those who produce or disseminate news, such as journalists and bloggers, as well as the media organisations themselves.

- News consumers are targeted by providing tools which help them to better understand the nature and sources of content (e.g. authentication of, and explicit metadata about, images and videos, better presentation of temporal information such as publishing dates, etc.). This also includes alerting news consumers to media entities who do not meet accepted standards - for example, having a history of suspicious funding, publishing fake material, or other dubious practices.

- News providers are targeted by providing them with methods and tools which can be implemented on their platforms directly, for example through the addition of metadata denoting sources and other credibility-related information for the stories and images/videos they provide, as well as better presentation of credibility-related indicators.

- Journalists are targeted by providing them with tools to help understand the nature of articles and media better (e.g. with provenance and credibility issues).

- Bloggers and citizen journalists, as well as media companies, are targeted by providing good practices and standards which all can follow when producing material (e.g. the Journalism Trust Initiative).

- Internet communications companies are also targeted by tools such as Newsguard. These are seen as a market for services that can help them to recognise purveyors of disinformation, and serve their audiences with these tools.

According to a Pew Research Center study (Mitchell, Gottfried et al., 2019), American news consumers expect the problem of misinformation in the news to be fixed by the news industry, even though they believe it is primarily the politicians who create it. Empowerment and credibility labelling responses put the onus on the consumer (and

sometimes the platform or news media organisations) to filter and interpret the content they encounter. Various research has indicated the potential value of explaining why something might be true/false and providing alternative factual information or a detailed explanation of why information is false (Ecker & Lewandowsky, 2010; Swire & Ecker, 2018). The assumption is that when people have aids for this, in the form of reliability signals, they will be more sceptical in the face of disinformation.

These responses thus aim to facilitate this task for the end user by providing mechanisms to signal disinformation. Changes in the way the news is presented to consumers, for instance, can be used to make the audience more aware of potentially misleading content. In particular, solutions are being proposed for protocols by which informative metadata can be added, made more visible, or even accompany published media wherever it is displayed (e.g. when further shared on social media or in the results of web search), as discussed in the following sections describing such initiatives.

Trust-based initiatives, which focus on highlighting good practices in the media, and promoting and rewarding compliance with professional standards, are based on the idea that particular media sources can be flagged as trustworthy, thereby encouraging common standards and benchmarks which can be adopted by all who produce journalistic content. Ultimately, adopting these standards could pave the way towards processes of certifications. Formal, or even informal, certification could lead to better discriminatory practices by consumers, but also to the adoption of better practices by media producers. An important outcome of trust initiatives is to build the faith of users in the media and counter their fears about the reliability of the content.

The theory of change represented by these initiatives can be summarised as follows:

- Relevant causes of disinformation:
  - news providers or internet communications companies spread disinformation, either because they are not trustworthy themselves and/or because they do not recognise it;
  - users do not recognise it, are influenced by it and also spread it further.
- Actions that the initiatives pursue to address the causes:
  - accrediting trustworthy news sources (and by implication, discrediting untrustworthy ones);
  - developing tools to empower media, internet companies and especially users to make better decisions about which information (and which media sources) can be trusted, as well as signposting issues to journalists and investigators;
  - developing protocols for providing better provenance information and making users aware of the importance of the source of content;
  - developing controlled-capture applications enables creators and distributors of images to create trust in their content.

- Desired outcomes of the initiatives:
  - better discriminatory practices by users;
  - adoption of better practices by news media and internet companies;
  - untrustworthy media sources are called to account;
  - faked media (video, images) become easier to spot and are less easily disseminated;
  - improved understanding by the public of disinformation and its playbook, etc.
- Potential impact of the initiatives:
  - confidence in the media and countering of fears about the reliability of information, leading to improved trust in place of a cynical relativism;
  - reduced rationales for producers of disinformation and encouragement of low-standard media to become more trustworthy;
  - increased spread of accurate information and reduced spread of inaccurate information;
  - increased agency for creators/distributors to assert trustworthiness, and for the users to assess it.

7.3.3 What output do empowerment and credibility labelling responses publish?

These initiatives publish a number of different kinds of output aiming to assist actors, ranging from general information, through methods and protocols, and sometimes even actual tools. These can be summarised as:

- provenance information of source material, and protocols for providing this;
- tools and resources for assessing credibility of news sources, feeding into accreditation schemes and content curation systems;
- methods and protocols for better provision of information to the end user, enabling improved awareness of trustworthy and untrustworthy information and sources;
- tools for rating news sources, articles and authors - carried out either by trained professionals (eg Newsguard) or community-driven (e.g. Credder).

We look at each of these in more detail below.

7.3.3.1 Provenance-tracking initiatives

Provenance-tracking initiatives emanate from a number of sources, and can be divided into three main subgroups: tools from news providers, tools at point-of-capture of images/video, and platform responses.

(**i**) **Tools from news providers** aim to assist news consumers to be more aware of misleading information, rather than try to prevent it occurring in the first place. For example, the News Provenance project aims to help users to better understand the nature and sources of content (e.g. authentication of and explicit metadata about images and videos, better presentation of temporal information such as publishing dates, etc.).

(**ii**) **Tools at point-of-capture of images/video** aim to track enhanced metadata and provenance, and confirm whether images and videos have been altered or not. For example, TruePic[448] is a venture-backed startup which is planning to work with hardware manufacturers (currently, just Qualcomm) to log photos and videos the instant that they are captured. Serelay Trusted Media Capture[449] also enables mobile phones to capture images and videos that are verifiable and for authenticity to be later queried by other apps. Amber[450] produces two tools: Amber Authenticate fingerprints recordings at their source, and tracks their provenance until playback, while Amber Detect uses signal processing and artificial intelligence to identify altered audio and video files. Eyewitness to Atrocities[451] is an app for mobile cameras which was developed for the specific purpose of documenting international crimes such that the footage can be authenticated for use in investigations or trials. Similarly to the others, it automatically records and stores metadata about the time and location of the recording, and includes a traceable chain of custody. All these (and other) tools are discussed in more detail in the Witness report (Witness Media Lab, 2019).

(**iii**) **Platform responses** come directly from the Internet communications companies themselves, e.g. image and video platforms. Some of these encourage users to add information to clarify that content complies with company standards and should not be removed. YouTube highlights the importance of adding context[452], for example to explain why graphic images might be necessary in newsworthy videos (and thus to prevent them being automatically rejected by YouTube in case they get flagged as being dubious). The social video company gives the example of a voice-over narration about the history of a protest - this kind of information is useful in helping a user to understand the provenance of a video.

Other kinds of responses involve directly labelling content; for example, YouTube does this to provide information on videos that highlight conspiracy theories (e.g. around the 1969 Apollo moon landing); or to indicate that content is from a state-funded broadcaster. Other platforms take similar action around inaccurate information on vaccinations, while in August 2020, WhatsApp introduced a feature which signals messages that have been forwarded five times or more, as an indicator of potential viral information.[453] Clicking

448 https://truepic.com/
449 https://www.serelay.com/
450 https://ambervideo.co/
451 https://www.eyewitness.global/
452 https://support.google.com/youtube/answer/6345162?hl=en
453 https://www.theguardian.com/technology/2020/aug/04/whatsapp-launches-factcheck-feature-aimed-at-viral-messages

on the magnifying glass symbol that automatically appears next to such a message initiates an online checking process which aims to reveal any known conspiracy theory or disinformation associated with the content of that message.

In early 2018, YouTube began labelling content in terms of whether the source counted as "state-funded media" in the company's definition.[454] In June 2020 Facebook introduced a similar policy, explaining that it was to help people understand whether the news they read is "coming from a publication that may be under the influence of a government".[455] Twitter introduced the practice some months later.[456]

7.3.3.2 Trust- and accreditation-based initiatives

Trust- and accreditation-based initiatives aim to develop and implement an agreed set of trust and transparency standards for media sources. These standards encompass transparency of media ownership and sources of revenues, as well as journalistic methods and the compliance with ethical norms and independence. Some of them aim to lead to a system of formal accreditation. Examples include:

- the **Journalism Trust Initiative**[457] (which involves Reporters Without Borders and its partners Agence France Presse, and the European Broadcasting Union);
- the **Trust Project**[458] (a consortium of top news companies, including the German news agency dpa, *The Economist*, The Globe and Mail, Hearst Television, the *Independent Journal Review*, Haymarket Media, Institute for Nonprofit News, Italy's *La Repubblica* and *La Stampa*, Reach Plc, and *The Washington Post*, and supported externally by various social media companies and search engines);
- the **Trusted News** initiative[459] set up by the BBC, which is planning a number of collaborative actions such as a rapid-response early warning system so that media (and other) organisations can alert each other rapidly in the case of disinformation which threatens human life. It is particularly tailored towards preventing the disruption of democracy during elections, with other actions based specifically around voter information and media education.

Trust-based initiatives also involve the development of tools and mechanisms for users to rate not only sources, but in some cases also individual articles, and/or journalists in terms of their credibility and trustworthiness. For example, **Credder**[460], which styles itself as "the trusted review site for news media" believes that "news should compete for trust, not clicks". It allows journalists and the public to review articles, measuring trust in not only the articles themselves, but also in the sources cited, and in the authors, and collates statistics on these. More generally, these tools use assessments and scoring of source quality (based on metrics such as accuracy and objectivity) to guide users towards higher-quality information and to help them to better discern and ignore low-quality information.

454 https://money.cnn.com/2018/02/02/media/youtube-state-funded-media-label/index.html
455 https://about.fb.com/news/2020/06/labeling-state-controlled-media/
456 https://www.bbc.co.uk/news/technology-53681021
457 https://jti-rsf.org/
458 https://thetrustproject.org/
459 https://www.bbc.co.uk/mediacentre/latestnews/2019/disinformation
460 http://credder.com

Newsguard[461] offers a browser plugin which aims to rate news sites based on what it defines as good journalism practices, via a nutrition-label methodology. This gives the reader additional context for their news, and also warns advertisers who might be worried about their brand's reputation to avoid unreliable sites. Green-rated sites signal good practices, following basic standards of accuracy and accountability, while red sites signal those with a hidden agenda or which knowingly publish falsehoods or propaganda. Additionally, grey sites refer to Internet platforms, while orange sites indicate satire. A colour-coded icon is shown next to news links on search engines and social media feeds, so that people are informed before they even click on the link. Additional information about the site, such as why it received the rating, can be obtained by hovering the mouse over the icon and/or clicking a button for additional information.

Décodex[462] is a tool created by French newspaper *Le Monde* to help people verify information with respect to rumours, exaggerations, twisted truth, etc. The tool works in two ways: a search tool enabling a user to check the address of a site for more information (e.g. to find out if it is classified as a satirical site); and a browser extension which warns the user when they navigate to a website or a social media account which has been involved in spreading disinformation.

MediaBiasFactCheck[463] is a tool which enables users to check the political bias of a particular media source. In the U.S. in particular, the public's most commonly given reason for media sites not making a clear distinction between fact and fiction is bias, spin and agendas, according to a report by the Reuters Institute (Newman & Fletcher, 2017). According to the philosophy behind the tool, the least biased sites are supposed to be the most credible, with factual reporting and sources provided. Questionable sources, on the other hand, exhibit features such as extreme bias, use of loaded words (conveying strong emotion designed to sway the reader), promotion of propaganda, poor sourcing to credible sites, and a general lack of transparency. However, the methodology behind the approach is not transparent, and it has been criticised itself for its quality.[464] Furthermore, it is not clear that it is a good idea to have a single number telling us how biased a news source is, as the situation is often more complex than this, and any notion of bias requires a baseline.

Maldito Bulo[465] is a browser extension created by one of the main debunking websites in Spain, Maldita.es[466]. The plug-in warns the user who has installed it if the consulted website has already published disinformation and how many stories have been debunked in the domain name.

KnowNews[467] is a browser extension which aims to help users understand which news sites are trustworthy or credible. It is developed by Media Monitoring Africa, which is an independent non-profit organisation from South Africa that promotes media freedom, media quality and ethical journalism.[468] The browser extension automatically classifies news sites based on their credibility, rating sites as credible, "dodgy" or not rated. The tool focuses on the content itself, however, directly evaluating information such as the

461 https://www.newsguardtech.com/
462 https://www.lemonde.fr/verification/
463 https://mediabiasfactcheck.com/
464 https://rationalwiki.org/wiki/Media_Bias/Fact_Check
465 https://chrome.google.com/webstore/detail/maldito-bulo/bpancimhkhejliinianojlkbbajehfdl
466 https://maldita.es/
467 https://newstools.co.za/page/knownews
468 https://mediamonitoringafrica.org/

authenticity of a photo, and is developed in partnership with Facebook and Google, as well as a number of other organisations.

The Knight Foundation's Trust, Media and Democracy Initiative[469] is anchored by the Knight Commission on Trust, Media and Democracy, a panel of people promoting more informed and engaged communities. This non-partisan group provides funding for seven initiatives:

- Cortico data analytics to surface underrepresented voices;
- Duke Tech & Check Cooperative + Share the Facts Database;
- First Draft fact checking network;
- AP fact checking;
- Reynolds Journalism Institute journalist training program;
- Santa Clara University Trust Project trust indicators;
- Your Voice Ohio strengthening ties to local communities;

The **IPTC** (international Press Telecommunications Council - the global standards body of the news media) has been collaborating with several initiatives around trust and misinformation in the news industry since 2018. This mainly involves working with The Trust Project and the Journalism Trust Initiative from Reporters Without Borders, but also to some extent the Credibility Coalition, the Certified Content Coalition and others, with the aim of identifying all known means of expressing trust in news content.

In April 2020, the IPTC published a draft set of guidelines[470] which aim to enable a news agency to add their own trust information to any news items they distribute. These indicators can also be converted to a standard schema.org markup language that can be added to HTML pages and automatically processed by search engines, social media platforms and specialised tools such as the NewsGuard plugin. This then enables users to see the trust indicators and decide for themselves about the trustworthiness of a piece of news.

The aim of the guidelines is to encourage news publishers to use trust indicators to show why they think they can be trusted, rather than just showing a certification of trustworthiness. Readers should be encouraged to follow links to understand the issues better. Indicators include those connected with editorial policy (e.g. statements about disclosure and correction practices, diversity and naming of sources, ethics, and feedback policies); party-level indicators (e.g. lists of other work by the author or provider; awards won; topics of expertise); organisation-level indicators (e.g. staff diversity; founding date of organisation; size etc.); piece-of-work-level indicators (e.g. details about dateline, editor, fact-checking; corrections; provider); person-level indicators (details about the author of the article); and type-of-work indicator (e.g. whether it is satire or not; what kind of report it is; background information, and so on).

469 https://knightfoundation.org/topics/trust-media-and-democracy
470 https://iptc.org/news/public-draft-for-comment-expressing-trust-and-credibility-information-in-iptc-standards/

These guidelines follow the idea of user empowerment by enabling users to make their own decisions rather than following blindly what is suggested to them. It also makes it easier for both information producers and consumers to follow established protocols. However, a limiting factor of this kind of methodology is that the guidelines are quite complex, and it takes time and effort on the part of the user to develop a full understanding about trustworthiness, and then to assess how it matches up to the claims of the organisation or content at hand. This increase in mental effort therefore best suits those who are already of a discerning nature, rather than those most susceptible to disinformation.

Finally, there are some specific themed initiatives which focus on a particular kind of rumour or topic, such as the **Vaccine Confidence Project** (Larson, 2018). This focuses on early detection of rumours about vaccines in an attempt to prevent them gaining impetus, but is entirely manual. A team of international experts monitors news and social media, and also maintains the Vaccine Confidence Index based on tracking public attitudes to relevant issues. While this is primarily a fact-checking operation, the project undertakes related research on trust and risk in this context and is dedicated to building public confidence and mitigating risk in global health. By listening for early signals of public distrust and questioning and providing risk analysis and guidance, they aim to engage the public early and thereby pre-empt potential programme disruptions.

7.3.4 Who are the primary actors behind empowerment and credibility responses and who funds them?

The actors and their sources of funding for these kinds of initiatives are quite varied, ranging from news media, through social media and internet communications companies, through to non-profit monitoring organisations.

Trust and accreditation initiatives are typically funded by either media companies, who are working together to develop formal systems of accreditation, or by monitoring organisations such as Media Monitor Africa and the Journalism Trust Initiative. Media companies clearly have an interest in establishing trust in news sources, though this raises a number of moral dilemmas (see further discussion on this in section 7.3.7 below).

Provenance initiatives are also funded by a variety of sources. Platform responses are typically funded by the social media companies, such as YouTube and Twitter, while other tools are provided by news providers such as the *New York Times*' News Provenance Project. Tools at point-of-capture are often funded by image and video software companies such as Adobe, as well as emanating from dedicated startups such as TruePic[471], Amber[472], and Serelay[473], while the open-source apps typically come from non-profit organisations, e.g. Tella[474], funded by the non-profit organisation Horizontal, and Eyewitness to Atrocities, funded by the International Bar Association in London in partnership with LexisNexis. The Guardian Project, which produces the open-source app ProofMode[475], is funded by a variety of organisations and foundations, including Witness, Google and various governments.

471 https://truepic.com/
472 https://ambervideo.co/
473 https://www.serelay.com/
474 https://hzontal.org/tella/
475 https://guardianproject.info/apps/org.witness.proofmode/

7.3.5 Response Case Study: COVID-19 Disinformation

The COVID-19 pandemic has highlighted the need for ways to help the public become more aware about disinformation. While educational responses are a major type of intervention, with many organisations and governments producing guides to staying well informed about dubious information and rumours surrounding coronavirus, there are also a few specific instances of organisations producing or highlighting credibility labelling and empowerment mechanisms. Many internet communications companies conferred prominent status on sources of reliable information on their services, such as the World Health Organisation and national health ministries. Signposting typically involves providing links to trustworthy sources of information, rather than explicitly pointing to untrustworthy sources. Examples of these efforts include the Harvard Medical School[476], which lists reliable sources of information on corona virus and provides tips on spotting this kind of knowledge resource.

An interesting method for assisting with the flagging of credibility comes from Wikipedia via WikiProjectMedicine[477], a collection of around 35,000 articles monitored by nearly 150 expert editors. Once an article has been flagged as relating to medicine, it becomes scrutinised more closely.[478] In this way, Wikipedia in some sense acts as a role model by having separate standards and stricter rules for particular situations (in this case, public health). Wikipedia also maintains an up-to-date page listing a variety of "misinformation" (as they term it) specifically about the coronavirus.[479]

Complementing these examples are initiatives to flag which content is dubious, thus indirectly also trying to help people to better understand which sources on the pandemic are genuine and provide verified information. (Content deemed dangerous to public and individual health is typically removed when identified). Many guidelines provided typically offer tips to users on not only how to spot a reliable source but also an unreliable ones), and they often offer advice on sharing (as has also been discussed in the first part of this chapter on MILresponses). One initiative to flag disinformation sources around COVID-19 comes from NewsGuard, who have partnered with BT to launch an online toolkit[480] to raise awareness of NewsGuard's online browser plugin, to help the UK public critically assess any information related to the global pandemic they come across online. The initiative is also backed by the UK Department for Culture, Media and Sport (DCMS) and the UK's library association. NewsGuard also made their browser plugin free until the end of July[481], specifically in the light of coronavirus. Previously it was available only as a subscription service, except to users of Microsoft Edge mobile devices. Since 14 May 2020 they have also extended this to all Microsoft Edge users on both mobiles and desktop applications, provided that the extension is used on that brower and downloaded in Microsoft Edge's store. They also set up a Coronavirus Misinformation Tracking Center[482] which signals all the news and information sites in the U.S., the UK, France, Italy, and Germany that they have identified as publishing materially false information about the

476 https://www.health.harvard.edu/blog/be-careful-where-you-get-your-news-about-coronavirus-2020020118801.
477 https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine
478 https://www.wired.com/story/how-wikipedia-prevents-spread-coronavirus-misinformation/
479 https://en.wikipedia.org/wiki/Misinformation_related_to_the_2019%E2%80%9320_coronavirus_pandemic
480 https://www.newsguardtech.com/press/newsguard-partners-with-dcms-and-bt-to-help-counter-spread-of-covid-19-fake-news-as-misinformation-peaks/
481 https://www.zdnet.com/article/newsguard-drops-its-paywall-to-combat-coronavirus-information/
482 https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/

virus.[483] The list includes sites that are notorious for publishing false health content, as well as political sites whose embrace of conspiracy theories extends well beyond politics.

While most credibility labelling responses in the fight against disinformation are still in their infancy due to ethical, legal and technological issues that are not yet fully solved, nevertheless the enormous amount of disinformation around corona virus and its potential seriousness is likely to become a strong driving force towards more effort in developing tools to aid users to supplement their existing media and information literacy levels as they navigate the "infodemic".

7.3.6 How are empowerment and credibility labelling responses evaluated?

Many of these initiatives are highly collaborative in origin and development, and are thus driven and evaluated through community efforts and advisory boards.

- Provenance-based initiatives are largely evaluated in-house. For example, the News Provenance Project is conducting user research to test the effectiveness of their proposed approach, in order to try to discover whether increasing access to metadata and supporting information helps consumers to better understand the veracity of professionally produced photojournalism.

- Accreditation-based initiatives are developing community-based standards with a solid background. For example, the Journalism Trust Initiative involves a very large number of advisory organisations, including official standards bodies such as the French Standardisation Association (AFNOR) and its German equivalent, the German Institute for Standardisation (DIN), as well as the European Committee for Standardisation (CEN).

- Trust-based initiatives are not always formally evaluated, but rely on community-driven input. For example, sites like Credder display the number (and content) of reviews submitted by users, so it is easy to derive statistical information about their usage and about agreement levels. What is less clear, however, is how helpful these reviews are to others. In other words, the quality of the input (the reviews themselves) can easily be judged and collective trust can be assessed, but the usefulness of the output and its overall impact is less easy to understand. Other sites, such as Newsguard, use trained analysts who are experienced journalists, to research online news brands in order to provide the ratings for sites. The lack of formal and independent evaluation for sites could be a major pitfall for such initiatives, especially if quality is dubious and users are unaware (Wilner, 2018: Funke & Mantzarlis, 2018b).

[483] At the end of March 2020, there were 144 sites, though this number is constantly growing.

7.3.7 Challenges and opportunities

There are a number of challenges arising from these kinds of initiatives, in particular the accreditation-based ones. Potential issues around authenticity and labelling are discussed in much fuller detail in the Witness report (Witness Media Lab, 2019), but we give a brief summary below. It should be noted first of all that the current overall impact of these initiatives is quite low since they are not yet widespread, as can be seen by the relatively low number of downloads of many extensions, Even NewsGuard with 78,000 users worldwide is like a drop in the ocean, especially since it is aimed at the general public.

The most important challenges are tackling impartiality, diversity and exclusion, in terms of who makes decisions about credibility and trust, and how they do it. There are also challenges to monitor and renew/revoke accreditation over a time period. Further, as accreditation and trust-based tools are implemented more widely, they become the de facto statements of trust across diverse media environments, resulting in inadvertent exclusion and inclusion of certain media entities. If authentication becomes the default for online media, this has an impact on pluralism as part of freedom of expression, in particular for those who are already disadvantaged in this respect. In particular, it will be a problem for those without access to verification technology or who may not wish to release potentially sensitive information such as their location.[484] This includes those who are in the Global South, use a jailbroken phone[485] , and who are also more likely to be women and live in rural areas. There are also potential problems of weaponisation of authenticity and provenance-based measures where their usage becomes obligatory, particularly in the context of 'fake news' legislation.

Furthermore, if credibility labelling is carried out by companies, there are risks in having a commercial organisation determining what is partisan and what is not, and non-transparent decisions and strategic biases could easily be incorporated. It also risks becoming a "one size fits all" approach, insensitive to cultural and societal specifics of particular countries, and implying that some fact-based narratives are intrinsically more worthy than other fact-based narratives – rather than signalling non-fact based content. Impartiality in such initiatives is often hard to maintain, and determining this on an ongoing basis can be problematic, which also raises questions about periodicity and mechanisms for the continuous review of labelling. Labelling initiatives can also impact negatively on the legitimate diversity of interpretation of narratives that are nevertheless fact-based.

A related issue is the overall accuracy and transparency of labelling tools. The science of provenance and related media forensics is not simple and not easily explained, so that while labelling information sources with their provenance information looks to be a simple solution, it is not only difficult and prone to error, but also not always obvious when it is incorrect. Credibility labelling is also potentially subjective if not opaque to the reader, as already discussed, and also can be error-prone whether manually or automatically carried out. The issues of accuracy around MediaBiasFactCheck have already been discussed earlier in this chapter, and they are certainly not the only tool with debatable quality of results. Who, then, should oversee the quality of such tools?

484 https://www.axios.com/deepfake-authentication-privacy-5fa05902-41eb-40a7-8850-5450bcad0475.html

485 A jailbroken phone is one which has been hacked to free it from the manufacturer's restrictions, and therefore has implications for the software, tracking options, etc. which can be used on it.

Moving on from reliability, we come to the question of psychological responses to the various mechanisms, in particular concerning the idea of empowerment. When content is marked or labelled in some way, there are a number of risks around the idea of interpretation. First, if a large number of false positives are flagged (i.e. if many legitimate or factual pieces of information are indicated to be suspicious or untrustworthy), people become habituated to this and have a tendency to over-interpret false alarms, believing that the tools are just over-sensitive and the labels are not believable. On the other hand, if content is not labelled with provenance or credibility (for example, if it is not clear how it should be labelled, and if false positives are to be avoided), then the assumption might be that the content is trustworthy, which is also potentially dangerous (see Pennycook et al., 2020). In terms of providing users with information, such as explanations around labelling, or even just explanations which help to empower the user in their decision-making, there is a tradeoff between providing sufficient information in enough detail to be clear, and in introducing too much complexity, which perpetuates further the divide between expert and consumer.

More generally, many of these initiatives are still quite young, and there is no broad adoption of any of these credibility labelling, provenance or controlled capture approaches. On the other hand, as discussed above, widespread use of such tools may lead to problems of strategic bias, exclusion, and unintended psychological perceptions. Fundamental questions also arise around how the approaches will be rolled out on a wider scale: for example, whether this will be in collaboration with platforms or in an alternative system. This relates in particular to technologies such as the use of blockchain, and there is a limited application of these approaches outside media and institutions in the Global North.

For content authentication systems at scale, there are issues in how to manage the challenges of doing this across technical and societal implications. For provenance tracking, there are a number of questions around the legitimate privacy and anonymity reasons, such as why people choose not to opt-in, or to only opt in for selected items of content, as well as technical constraints. This leads to the question of how to ensure that trust is on an opt-in rather than an obligation basis, and thereby only a signal of trust, rather than a confirmation. This latter is an important dilemma for many other kinds of anti-disinformation initiatives - in order to be effective, these mechanisms need to be widespread, but this causes serious problems when - sometimes for legitimate privacy reasons such as whistle-blowing - people do not choose to authenticate their data, or when relevant verificatory information is incorrect or missing.

Finally, in terms of user empowerment, there are a number of questions around best practices for managing and presenting complex information. As discussed above, information needs to be presented in a simple yet still meaningful way in order for the general public to be able to make appropriate use of it and understand its implications, but too simple a presentation may lead to misinterpretation by suggesting that issues of verification and trust are black and white. On the other hand, in order not to overwhelm the user, systems of progressive disclosure (by means of breaking down detailed relevant information related to trust and credibility into deeper levels to be explored for further understanding) could be a suitable approach, but have not yet been adopted. Clearly, there is no one-size-fits-all solution to the problem of empowering users to

become more discerning about the information they consume, and further social and psychological research is still very much a requirement, as well as the technical and legal issues to be resolved.

Another challenge linked to the psychological issue is the tradeoff between the empowerment of the user by providing them with pointers to helpful information about provenance and trust, thereby avoiding external bias and simultaneously helping to educate the user, and the fact that the onus is now on the users to make the decisions, when they still may not be sufficiently equipped to interpret the results correctly. The interpretation of labels adds a significant additional neural processing load for the consumer, a known factor in both the spread of disinformation and in issues of unconscious bias and filter bubbles. Conversely, these mechanisms also provide a heuristic shortcut that may not be accurate (see for example a recent report by Witness discussing the history of verified checkmarks, and how they default to erroneous instinctive rather than rational thinking) (Witness Media Lab, 2019). Tools and practices which allow a consumer to verify that a particular piece of content came from a particular source also do not help if the consumer does not properly understand the reliability of that source. Thus a holistic approach that incorporates both aspects, and educates the user to use proper discernment in their news consumption, is still critical.

On the opportunities side, many of the challenges listed above can be addressed through transparency, consultation and respect for pluralism and diversity within freedom of expression. Further, one of the greatest strengths of empowerment and especially credibility labelling responses is that the indicators produced are easy to interpret with little training required. For example, 'traffic light' systems make it very clear what is trustworthy and what is suspicious. This is particularly important for the general public who cannot be expected to become expertly media and information literate overnight, despite the benefits that educational initiatives afford, as discussed in the previous chapter. Nor can the general public be expected to fact-check all the content they come across. Thus, the aids discussed above supplement what skills consumers themselves bring to negotiating with content.

These systems can also lead to long-term benefits such as news providers becoming more trustworthy overall, because when their failings are highlighted compared with certified performers, there is greater incentive to improve. Taking this further, a widespread adoption of sets of certifiable standards for the media industry also has potential benefits, such as helping to strengthen the economic situation of legitimate publishers

Additionally, provenance-tracking initiatives which help the user understand the source and nature of the material, or in some way verify its content, save time. This is important to journalists in the fast-paced media world, but also to ordinary members of the public who do not want to spend a lot of effort in checking sources, even if they understand the importance of it. Along with the trust and credibility tools, this time-saving feature in turn helps to drive the adoption of good practices and standards not only by large media companies, but by all who produce media content, such as bloggers and citizen journalists. Finally, if such solutions are successful, they can be adopted by media organisations more broadly. For example, blockchain-based protocols can be used to share metadata along with media content wherever that content goes.

7.3.8 Recommendations for empowerment and credibility labelling responses

In general, the use of consumer aids entailing standardisation and certification (without compromising pluralism and diversity), as well as approaches that can be rolled out globally and across different platforms, can be encouraged.

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

Internet communications companies and news media could:

- With full respect for media pluralism and diversity, adopt certifiable standards with respect to credibility labelling of news institutions.
- Consider clear and simple, time-saving content labelling approaches, with full transparency about the criteria involved, the implementation process at work, and independent appeal opportunities.
- Avoid quick fix solutions, which can be misleading and have unwanted consequences, such as leading people to blindly trust flags and indicators which may not tell the whole story – or leading to people discounting these signals due to 'false positives' or bias.
- Experiment with signposts and indicators which encourage people to think for themselves, and raise the level of their critical Media and Information Literacy.
- Ensure that empowerment and labelling responses operate in tandem with educational responses for best effect.
- Implement better mechanisms for assuring transparency and accountability of institutions and communities engaged in the design and implementation of empowerment and credibility labelling approaches, as well as their independent evaluation.
- Develop credibility responses with great care, especially with consideration towards less developed countries, smaller media and technology companies, and disadvantaged communities who could be negatively affected by inflexible solutions that are insensitive to inequalities and media pluralism and diversity.

Researchers and civil society could:

- Experiment with the implementation and adoption of global solutions (such as blockchain protocols) for provenance tracking and avoid piecemeal approaches.
- Track practices within the media and internet communications companies as a whole, including assessing the significance of metadata for content no matter where that content ends up.

This study presents an original typology of responses to disinformation which addresses the entire spectrum of responses on a global level, capturing a multitude of initiatives and actors. Moreover, the research offers a unique approach: it places freedom of expression-related challenges and opportunities at the core of the analysis.

Particularly novel, is the study's emphasis on identification and explication of **11 different types of disinformation responses** assessed in terms of the objects they focus upon, instead of framing them through a lens on the key actors involved. Similarly, there is the global scope of the project - with many initiatives included from the developing world to ensure geographical diversity.

Additionally, the diverse nationalities and disciplines of the researchers associated with this project allowed a multiplicity of perspectives to emerge and converge, producing a rich and substantial piece of policy research which is tied to both practice and impact, emphasising technological measures, State interventions, pedagogical initiatives, state and journalistic interventions.

Finally, there is an attempt to deconstruct disinformation in a fresh way, by investigating the underpinnings of these responses in terms of the implied theories of change behind them, as well as an analysis of their targets, and the funding sources they depend upon.