Photo credits: Shutterstock, igorstevanovic

Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

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

Executive Summary

This content is part of the wider study: "Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression", a Broadband Commission research report available at: https://en.unesco.org/publications/balanceact

Editors:

Kalina Bontcheva & Julie Posetti

Contributing authors:

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

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

ISBN 978-92-3-100403-2

Executive Summary

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

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

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

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

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

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

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

1. Instigators:

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

2. Agents:

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

3. Messages:

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

4. Intermediaries:

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

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

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

5. Targets/Interpreters:

- Who is affected? Are the targets individuals, journalists and scientists, systems (e.g. electoral processes; public health; international norms); communities; institutions (like research centres); or organisations (including news media);
- What is their online response and/or real-life action? This question covers responses such as inaction, sharing as de facto endorsement, liking, or sharing to debunk disinformation. Is there uncritical news reporting (which then risks converting the role of a complicit journalist/news organisation from target into a disinformation agent)?
- Responses identifying what messages count as disinformation, investigating the instigators and agents, identifying the intentions and targets;
- Responses aimed at circumscribing and countering the agents and instigators of disinformation campaigns;
- Responses aimed at curtailing the production and distribution of disinformation and related behaviours, implemented particularly by intermediaries and media;
- Responses aimed at supporting the targets/interpreters of disinformation campaigns.

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

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

   i. Monitoring and fact-checking

   ii. Investigative

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

   iii. Legislative, pre-legislative, and policy responses

   iv. National and international counter disinformation campaigns

   v. Electoral responses

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

   vi. Curatorial responses

   vii. Technical and algorithmic responses

   viii. Demonetisation responses

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

   ix. Normative and ethical

   x. Educational

   xi. Empowerment and credibility labelling responses

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

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

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

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

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

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

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

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

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

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

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

Individual states could:

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

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

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

Political parties and other political actors could:

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

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

Electoral regulatory bodies and national authorities could:

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

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

Law enforcement agencies and the judiciary could:

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

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

Internet communications companies could:

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

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

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

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

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

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

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

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

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

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

Assessment Framework for Disinformation Responses

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

#DISINFODEMIC
#THINKBEFORESHARING
#SPREADKNOWLEDGE

BROADBAND COMMISSION
FOR SUSTAINABLE DEVELOPMENT

9 789231 004032