2 Typology of Disinformation Responses

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

Chapter 2 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

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

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

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

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

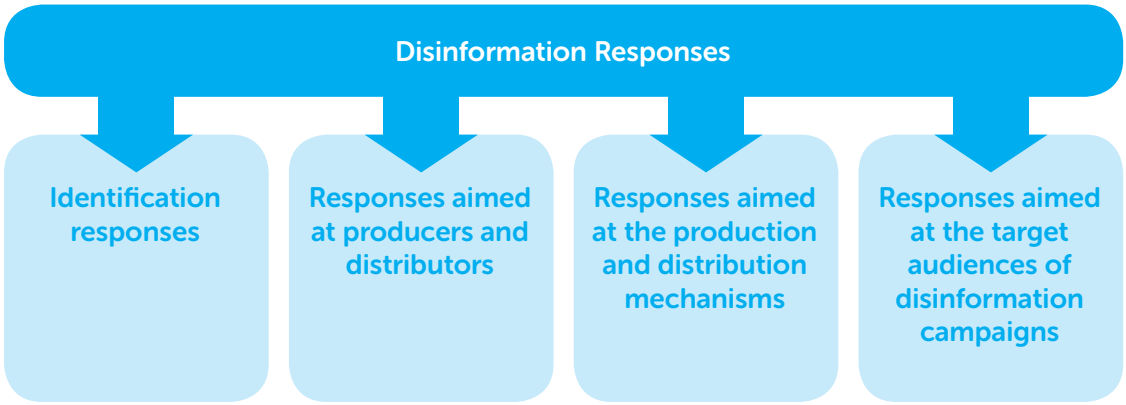

Figure 1. *Top-level categories of disinformation responses*

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

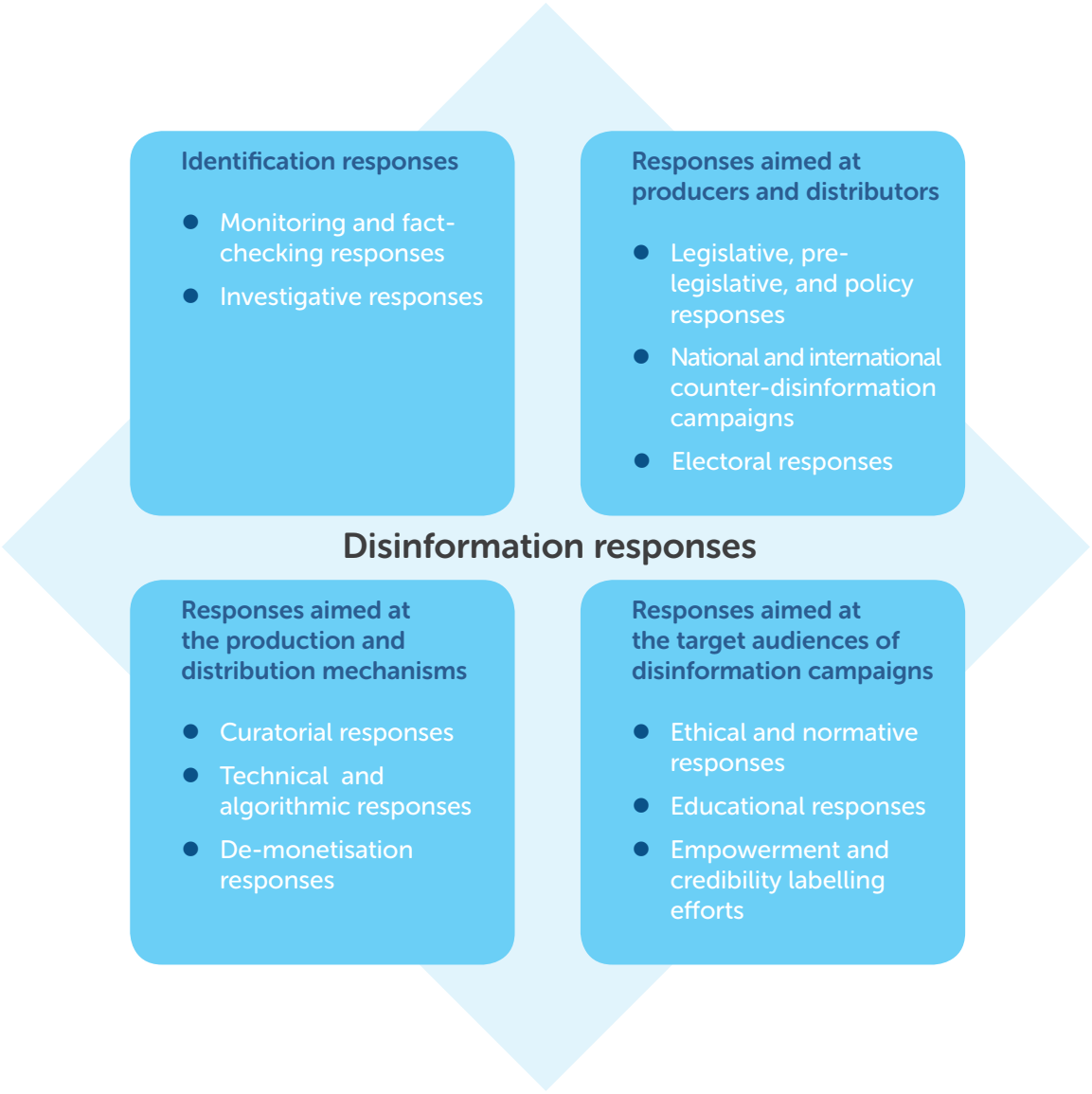

Figure 2. *The 4 top-level response categories and their eleven sub-categories.*

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

- **Monitoring and fact-checking responses**, which tend to be carried out by news organisations, internet communications companies, academia, civil society organisations, and independent fact-checking organisations, as well as (where these exist) partnerships between several such organisations.
- **Investigative responses**, which go beyond the question of whether a given message/content is (partially) false, to provide insights into disinformation campaigns, including the originating actors, degree of spread, and affected communities.

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

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

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

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

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

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

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

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

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

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