VUB

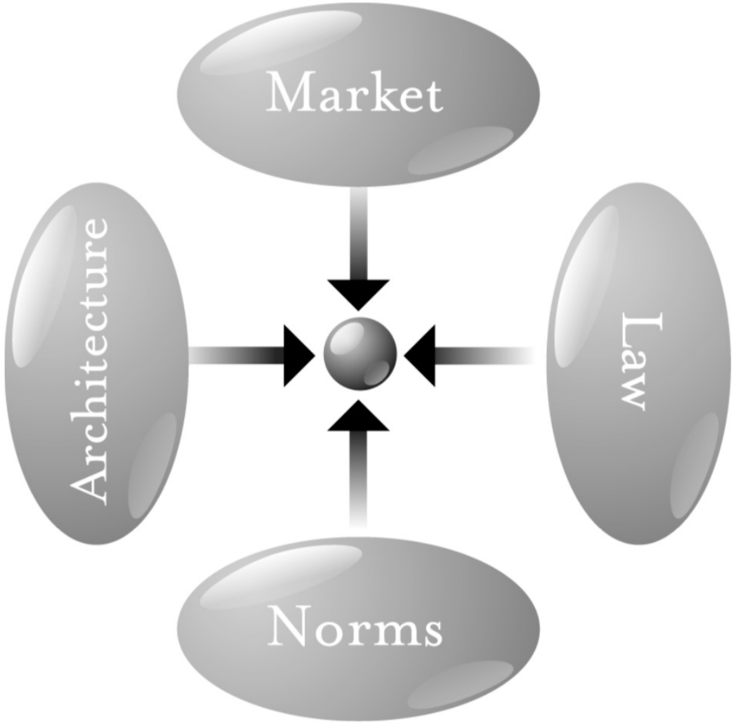
BRUSSELS SCHOOL OF GOVERNANCE

CHALLENGES AND OPPORTUNITIES OF PLATFORM RESPONSES TO DISINFORMATION FOR FREEDOM OF EXPRESSION AND THE MEDIA

OSCE EXPERT MEETING: INTERNATIONAL LAW AND POLICY ON DISINFORMATION IN THE CONTEXT OF FREEDOM OF THE MEDIA

Trisha Meyer, 14 May 2021

European Pillar of Social Rights
Socle européen des droits sociaux
Europese pijler van sociale rechten
ec.europa.eu



Are we discussing regulation through technology or regulation of technology (Lessig 2006)?

What are the opportunities and challenges of different modes of regulation?

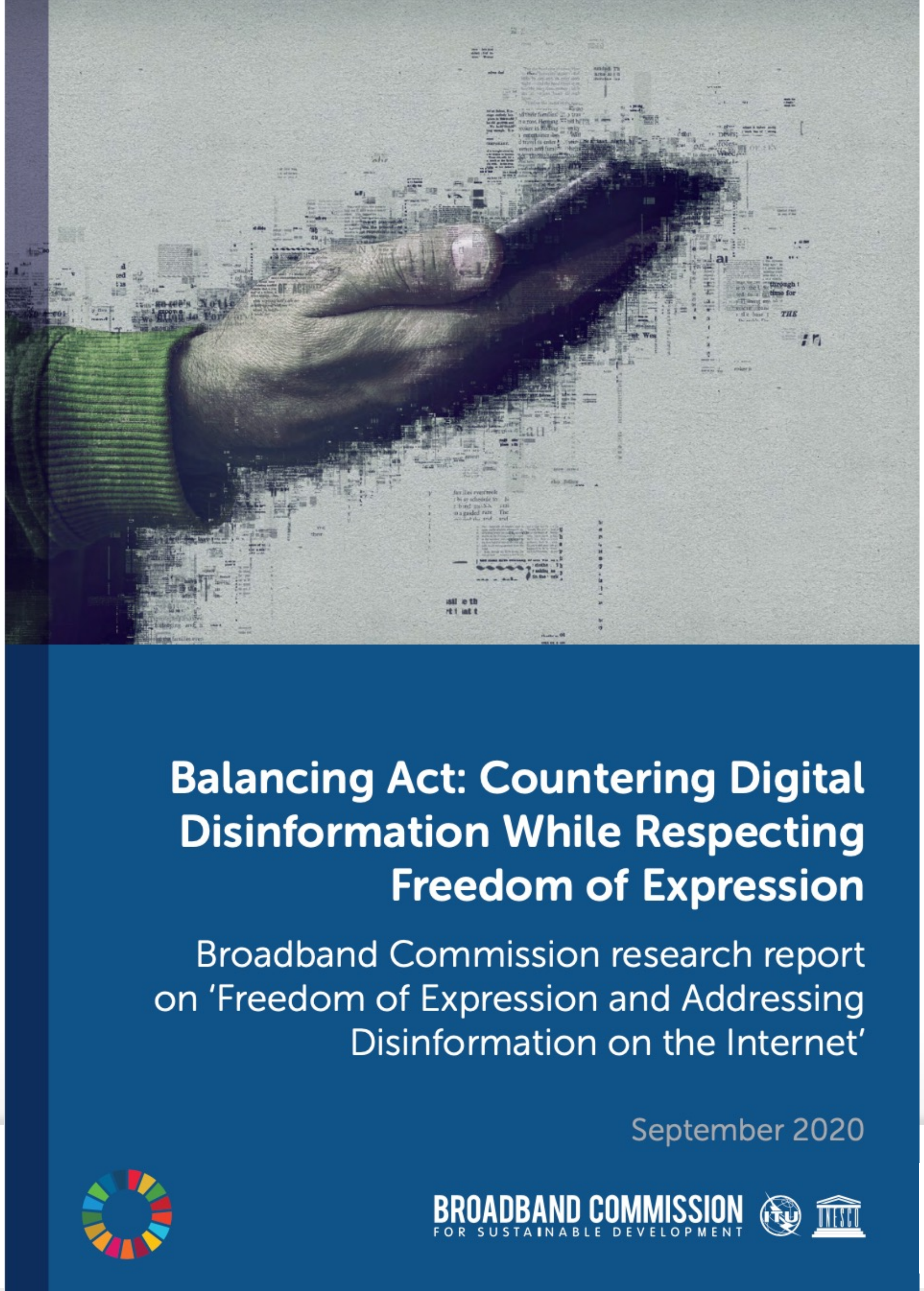
FREEDOM OF EXPRESSION

Article 19 of the UN Declaration of Human Rights and echoed in the International Covenant on Civic and Political Rights:

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

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

Quid platform responses? Quid automated content moderation?

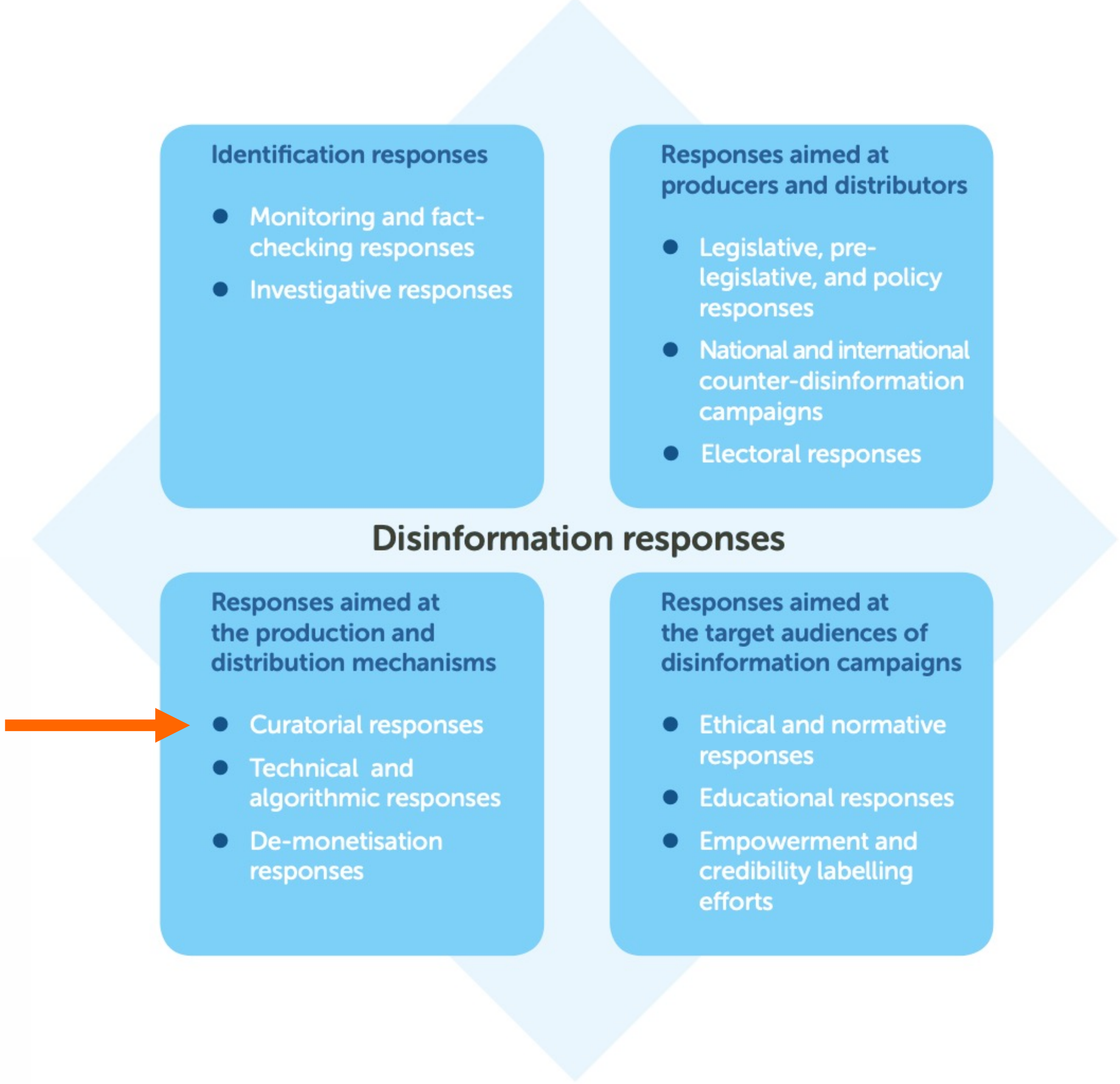


Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression

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

September 2020

BROADBAND COMMISSION FOR SUSTAINABLE DEVELOPMENT



Identification responses

- Monitoring and fact-checking responses
- Investigative responses

Responses aimed at producers and distributors

- Legislative, pre-legislative, and policy responses
- National and international counter-disinformation campaigns
- Electoral responses

Disinformation responses

Responses aimed at the production and distribution mechanisms

- Curatorial responses
- Technical and algorithmic responses
- De-monetisation responses

Responses aimed at the target audiences of disinformation campaigns

- Ethical and normative responses
- Educational responses
- Empowerment and credibility labelling efforts

Report available at: https://www.broadbandcommission.org/Documents/working-groups/FoE_Disinfo_Report.pdf

PLATFORM POLICY RESPONSES

CONTENT AND ACCOUNT MODERATION

- Continued analysis for EU Disinfolab COVID-19 Disinformation Research Hub
- Analysis of curatorial responses to disinformation outlined platform terms of service, community guidelines and editorial policies
- Responses to COVID-19 and US related disinformation have focused on
  - **promotion of authoritative sources**
  - **provision of free ad-credits**
  - **rapid expansion of disinformation policies to limit/remove content**
  - **increased automation of content moderation**
  - **but with (even) less guarantee of review and appeal**

Summary available at: https://www.disinfo.eu/publications/how-platforms-are-responding-to-the-disinfodemic AND https://www.disinfo.eu/publications/one-year-onward-platform-responses-to-covid-19-and-us-elections-disinformation-in-review/

FOCUS ON COVID-19 AND US ELECTION DISINFORMATION

Main response type per platform	Facebook	Google	TikTok	Twitter
Flagging/labelling content	E		C	C / E
Blocking/removing content	C / E	C	C / E	C / E
Limiting/demoting content	C / E		E	E
Prioritizing/amplifying content	C	C / E	C / E	C / E
Account-specific	E			E
Advertising-specific	C / E	C / E	C	C
User-specific			C	
Review/disinformation research-specific				

Comparison of platform responses to COVID-19 and US elections related disinformation (own compilation) [main responses related to C = COVID-19; E = US elections]

CHALLENGES AND OPPORTUNITIES

PRIVATE JUDGES AND ENFORCERS OF FREEDOM OF EXPRESSION

- In response to the ongoing health and political ‘disinfodemic’, online platforms took unprecedented measures to minimize harm through content (and account) moderation

- Some policy updates were clearly planned; others were kneejerk reactions – in any case, editorial policies tend to be stricter than legally required (at least in the jurisdiction of legal registration)

- Private companies with global reach are acting, in an uncoordinated manner, as definers, judges and enforcers of acceptable expression on their services

- In the absence of harmonised standards and definitions, each company uses its own ‘curatorial yardstick’, with no consistency in enforcement, transparency or appeal – in other words accountability, across platforms

RECOMMENDATIONS

TOWARDS HUMAN RIGHTS PRINCIPLES FOR CONTENT MODERATION (KAYE 2018)

- Platform responses to the ‘disinfodemic’, in particular increased automation, confirm the urgent need for transparent and **measurable content moderation policies**, that **support human rights** and are implemented equitably on a truly global scale
- Detailed and frequent (public) **transparency reports**, including specific information on spread of disinformation, suspension of accounts, removal of content and other steps against disinformation, including content promotion, demotion and demonetisation
- **Empowerment of users** through notice of action, ability to appeal decisions, including machine-driven ones, and access to journalism from independent and professional news organisations
- Multistakeholder **social media councils**, including independent researchers, journalists and civil society organisations, allowing for industry-wide scrutiny, complaints and recommendations of interpretation and implementation of standards