Introduction

1

Authors: Kalina Bontcheva and Julie Posetti

Chapter 1 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

This global study seeks to map and deepen understanding of diverse international responses to disinformation, along with the impacts of counter-disinformation measures on the right to freedom of opinion and expression, as described in Article 19 of the United Nations' Universal Declaration of Human Rights[1]:

> *"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."*

Freedom of expression rights, including press freedom and the right to access information, are upheld in tandem with privacy rights, which are also enshrined in international human rights law. So, where relevant, this study also touches on online privacy and dignity issues. Further, it situates the problem of disinformation in the context of the enabling role of the internet - especially the social web - in both improving access to information, and as a disinformation vector. It discusses in detail the potential for responses to disinformation to curb freedom of expression and suggests ways to avoid such impacts.

Although many studies and policy papers on disinformation have already been published by governments, international organisations, academics, and independent think tanks, this study offers novel contributions through its development of a systematic typology of the range of responses to disinformation which is applied internationally:

1. Addressing the entire spectrum of disinformation responses, rather than e.g. just educational or legal or technological responses;
2. Categorising responses according to the target of the intervention, rather than in terms of the means used or the actors involved;
3. Assessing responses in terms of key assumptions and significance from a freedom of expression point of view;
4. Representation of geographically diverse issues, cases and responses, including an emphasis on the Global South;
5. Providing an overview of disinformation responses aimed at 'flattening the curve' of the COVID-19 'disinfodemic' (Posetti & Bontcheva 2020a; Posetti & Bontcheva 2020b).

There are diverse definitions applied to false and misleading information, but for the purposes of this study the term disinformation is used throughout to broadly refer to content that is false and has potentially damaging impacts - for example, on the health and safety of individuals and the functionality of democracy. For many analysts, the intent of the agent producing or sharing the inaccurate content can also differentiate disinformation (deliberate falsehood) from misinformation (unconscious falsehood). This study accepts the role of such a distinction, which also implicates different types of remedies. Nevertheless, the impact of the false content, irrespective of intentions, can be the same. It is this focus on the potentially damaging effects of fabricated and misleading content, rather than the motivation for its creation and dissemination, that explains the broad use of the term disinformation here as the umbrella term - irrespective of intentionality or underlying behaviour in spreading such messages. This rationale is further explained in section 1.2 on definitions below.

[1] https://www.un.org/en/universal-declaration-human-rights/

Disinformation (as opposed to verifiable information) can cause harm since it may serve to confuse or manipulate citizens, create distrust in international norms, institutions or democratically agreed strategies, disrupt elections, or paint a false picture about key challenges such as climate change. It can also be deadly, as the COVID-19 'disinfodemic' has illustrated (Posetti and Bontcheva 2020a; Posetti and Bontcheva 2020b). Disinformation is typically organised by both state or non-state actors, including individuals and organised groups. It is created, spread and amplified both organically, by people who believe it, and artificially through campaigns that make use of technology such as bots and recommender algorithms. It is crafted to exploit cognitive biases such as attentional and confirmation biases, while using astroturfing[2] techniques to stimulate what is known as the 'bandwagon effect' (Schmitt-Beck, 2008), creating the impression of widely shared beliefs around a particular issue or item. Frequently, disinformation campaigns aim to target, discredit, and silence those who produce verified information or hold opposing views, including politicians, journalists, human rights campaigners, scientists, and others. Many disinformation agents carry out campaigns that are also networked across different platforms and combined with threats, intimidation and disruptive tactics.

In particular, disinformation negatively impacts citizens' rights to privacy, freedom of expression and access to information. In turn, however, many efforts to tackle online disinformation can also interfere with these fundamental human rights, as discussed throughout this report. Tools, measures and policies to address the disinformation problem therefore need to ensure that the rights of citizens are protected, and that their interests are represented. This means taking an approach that acknowledges how the issues affect stakeholders such as journalistic actors, civil society organisations, and the internet communications companies.[3] Frequently, however, these rights and interests are in tension in the struggle to identify, curtail and counter disinformation. For example, what's the interplay between content moderation, freedom of speech, and algorithmic amplification of misinformation?

Under human rights law, expression of false content - like other expression - is protected, with some exceptions. For example, under the International Covenant on Civil and Political Rights, certain forms of hate speech, incitement to violence, and speech that threatens human life (including dangerous health disinformation) can attract legitimate restrictions for reasons such as the protection of other human rights, or for public health purposes. Nevertheless, inasmuch as speech does not reach this threshold of legitimate restriction, people have a right to express ill-founded opinions and make non-factual and unsubstantiated statements - ranging from claims that "The earth is flat" to opinion like "The unusually cold weather we are experiencing means that global warming must be a myth in my view" - including on social media (Allan, 2018). On the other hand, falsehoods designed to defraud people financially, defame a person's reputation, or suppress voter turn-out, may be fairly penalised under criminal or civil law in many cases. All this makes tackling disinformation even more complex from the point of view of freedom of expression.

2 'Astroturfing' is a term derived from a brand of fake grass used to carpet outdoor surfaces to create the impression that it is natural grass cover. In the context of disinformation, it involves seeding and spreading false information, targeting audiences and journalists with an intention to redirect or mislead them, particularly in the form of 'evidence' of faux popular support for a person, idea or policy. See also Technopedia definition: https://www.techopedia.com/definition/13920/astroturfing

3 Throughout this report, the term 'internet communications companies' is used to refer to large companies in the sphere of search engines, social media sites and messaging apps. This avoids the practice of referring to these companies generically as 'the platforms' in order to underline their diversity, and because they are not neutral or passive technological infrastructural services but institutions with interests, obligations and configurations that have significant bearing on information, disinformation and communications.

Contemporary expression is closely intertwined with the combination of information technologies and internet communications companies which, coupled with growing broadband access, enable the instantaneous dissemination of information within global networks that are accessible to billions of people. This facilitates freedom of expression and potentially opens up a far wider range of viewpoints and information sources to citizens than ever before. In a world divided between information-rich and information-poor, this is seen as a boon to people who have previously been uninformed. Conversely, however, these tools of freedom of expression have been increasingly weaponised by actors seeking to manipulate public opinion by inserting and amplifying false and misleading content within the online information ecosystem.

The increasing availability of information, coupled with the potential for more diverse news diets, could widen the range of ideas to which people are exposed. Within the vast sea that is the contemporary information ecosystem, there are credible information providers like those journalism producers who do live up to the standards of independent professionalism, independent research institutes, other producers of reliable public interest information (e.g. reputable health advice providers), and well-informed commentators. But there is also a mass of other players with different standards of truthfulness, diverse ethics and varying motives creating a powerful rip current within this sea. Consequently, citizens can feel overwhelmed by the flood of content they are exposed to online, and they can come to rely on spurious sources that appeal to their biases and reinforce their pre-existing beliefs or identities. As a result, in place of being uninformed, they may become actively disinformed, or indirectly misinformed.

Recent research has demonstrated that disinformation affects different countries to various extents (Humprecht, Esser & Van Aelst, 2020). Increased ideological segregation and political polarisation are some of the key drivers behind the elevated production and spread of online disinformation in some countries (Humprecht, Esser & Van Aelst, 2020). By contrast, other research indicates that digital information consumption can lead to exposure to a broader range of information sources, although it does not necessarily follow that the content is itself more diverse, nor that the beliefs held are therefore diversified. However, repetitious exposure to falsehoods is known to reduce resistance to disinformation, as does exposure to high levels of populist communication (Humprecht, Esser & Van Aelst, 2020).

Conversely, resilience to disinformation is higher in countries where trust in news media is high and public service media provision is strong. Moreover, low public trust in news media and democratic institutions can lead to highly selective information consumption through online echo chambers that amplify disinformation and deepen polarisation.

Consequently, there is an urgent need to not only address disinformation, but also to take steps towards rebuilding the social contract and public trust in agreed international norms and standards: strengthen democratic institutions; promote social cohesion particularly in highly divided societies; and engage dialogue-building tactics to address entrenched groups and actors online.

This is why it is imperative to examine the diverse responses to disinformation globally, and to develop frameworks to help understand and assess these responses through a freedom of expression lens. That is the primary work of this study, research for which was conducted between September 2019 and July 2020.

Before this work turns to deciphering and dissecting these dimensions, it is necessary to outline the parameters for the research, explain the key terms used, and consider some examples of online disinformation, along with their relationship to propaganda, misinformation, and hate speech.

1.1 Techniques of online disinformation

The ubiquitous presence of online disinformation poses serious questions about the role of search, social media and social messaging and the internet more widely in contemporary democracies. Examples of digital disinformation abound, ranging from election interference to medical disinformation (e.g. vaccination[4]; coronavirus[5]) and these frequently involve threats of physical harm, privacy risks, and reputational damage to individuals and public health.

While disinformation is often studied in regard to Twitter, Facebook, and YouTube, it also exists on many other social platforms (e.g. Reddit, Instagram, TikTok, 4chan, Pinterest[6]), messaging apps (e.g. WhatsApp, Telegram, SnapChat, and iMessage), and internet search engines (e.g. Google). There are also dedicated disinformation sites (e.g. Infowars, Q-anon). Additionally, other actors and intermediaries (e.g. ISPs, cloud computing providers) will also be referenced here where relevant. The study, while comprehensive at the time of writing, also acknowledges the need to continue research into emerging disinformation mechanisms and new and rapidly evolving social platforms, including those received or perceived mainly as entertainment and social spaces (e.g. TikTok) and not as platforms for political and other purposes.

While political actors and States are often significant producers and conduits of disinformation (Brennan et al 2020; Billings 2020; Bradshaw & Howard 2019), the emphasis of this report is not on disinformation sources and actors, but on the responses to disinformation found across the world. Among these responses, States and political actors have a critical role to play in stemming disinformation at the source - including within their own 'houses'. Their relevance is especially assessed in regard to responses concerning counter-messaging, legislation and policy, elections and normative interventions.

Many mainstream news producers - online and offline - struggle to remain a reference point for those seeking trustworthy information within this wider ecology of communications. Through weak standards of verification, manipulation by outside actors, and even complicity (e.g hyper-partisan media), news outlets have also become vectors for disinformation in certain cases.

Nevertheless, the legitimating and agenda-setting public role of critical independent news media also makes them prime targets for purveyors of disinformation. In the case of orchestrated disinformation campaigns, attacks are frequently deployed against legitimate and authoritative information sources - such as credible news media and journalists - through hacking, disruption, and other tactics of intimidation and surveillance, with a view to a holistic strategy for advancing disinformation and wider objectives. Many orchestrated disinformation campaigns are State-initiated and/or connected to political and geopolitical actors, and this is relevant to understanding State roles in the responses to disinformation. However, the primary purpose of this report is to unpack the diverse modalities of response to the global disinformation crisis, rather than assessing the initiators and agents and their motives.

4 https://firstdraftnews.org/long-form-article/first-draft-case-study-understanding-the-impact-of-polio-vaccine-disinformation-in-pakistan/

5 https://www.poynter.org/fact-checking/2020/coronavirus-fact-checkers-from-30-countries-are-fighting-3-waves-of-misinformation/%20

6 https://medium.com/dfrlab/trudeaus-and-trudeaunts-memes-have-an-impact-during-canadian-elections-4c842574dedc

Within the ecosystem, disinformation knows no boundaries, but rather permeates multiple communication channels by design, or through redistribution and amplification fuelled by the architectures of interconnecting peer-to-peer and friend-to-friend networks.

With respect to types of content, three main disinformation formats have been identified for this study, based on the modality of the content (e.g. text, image, video, audio, mixed) and the way it has been constructed or manipulated:

1. **Emotive narrative constructs and memes**: False claims and textual narratives[7] which often (but not always) mix strong emotional language, lies and/or incomplete information, and personal opinions, along with elements of truth. These formats are particularly hard to uncover on closed messaging apps and they are applied to a range of content from fabricated 'news' to problematic political advertising.

    - **False/misleading narratives** emulating formats like news writing or documentary, and which typically mix false textual claims or incomplete information with personal opinions, along with images and/or video and/or audio, which themselves could be inauthentic, manipulated, or decontextualised. Appropriated content from other websites is sometimes used to create a misleading overall impression of being a neutral news-aggregator.

    - **Emotional narratives** with strong personal opinions, images and/or videos and audio, which may be inauthentic, manipulated, or decontextualised, and which also seek to dictate interpretations of particular information at hand, e.g. minimising its significance, smearing the source.

2. **Fraudulently altered**, **fabricated**, **or decontextualised images**, **videos[8] and synthetic audio[9]** used to create confusion and generalised distrust and/or evoke strong emotions through viral memes or false stories. These are also applied to a wide range of content from political propaganda to false advertising. Among these techniques we can distinguish:

    - **Decontextualised images and videos** that are unchanged or almost unchanged with high level of similarity, and often including copies that are used for clickbait purposes;

    - **Altered decontextualised audio**, **images and videos that** are cut in length to one or several fragments of the original audio or video, or changed to remove a timestamp in CCTV camera footage, for example. These are also called 'shallow fakes';

    - **Staged videos** e.g. produced on purpose by a video production company;

    - **Tampered images and videos** that are created with the help of editing software to remove, hide, duplicate or add some visual or audio content;

7 A database of over 6,000 fact-checked false claims and narratives on COVID-19 from over 60 countries: https://www.poynter.org/coronavirusfactsalliance/

8 Decontextualised images or videos are pre-existing, authentic content, which is re-purposed as part of a false narrative to spread disinformation, e.g. an old video of people praying was used in a far-right tweet claiming that Muslims are flouting social distancing rules.

9 See definition below

- **Computer-Generated Imagery** (**CGI**) including deepfakes (false images/videos generated by artificial intelligence) that are entirely computer-generated, or mixed with a blend of pre-existing image/footage/audio.

- **Synthetic audio**: Speech synthesis, where advanced software is used to create a model of someone's voice is a relatively new branch of deepfakes. This involves replicating a voice, which can verbalise text with the same cadence and intonation as the impersonated target. Some technologies (e.g. Modulate.ai) allow users to create completely synthetic voices that are able to mimic any gender or age. (Centre for Data Ethics and Innovation 2019)

3. **Fabricated websites and polluted datasets**, including false sources, manipulated datasets, and fake government or company websites (Trend Micro, 2020). This category also includes websites using names that make them sound like news-media and which publish seemingly plausible information in the genre of news stories, e.g. reporting bogus cases of COVID-19 (Thompson, 2020).

These different disinformation modalities are harnessed in a range of potentially harmful practices, including but not limited to:

- State-sponsored disinformation campaigns;
- (Anti-)Government /Other political propaganda;
- Political leaders generating and amplifying false and misleading content
- Clickbait[10];
- False or misleading advertisements e.g. connected to politics, job adverts;
- Impersonation of authoritative media, fact-checking organisations, people, governments (false websites and/or social media accounts, bots);
- Astroturfing campaigns;
- Fake products and reviews
- Anti-vaccine, coronavirus, and other other health, medical and well-being related misinformation;
- Gaslighting[11];
- Inauthentic identities and behaviours;

Overt satire and parody are excluded from this list of communication practices, even though in some instances these may have the potential to mislead and thus cause harm to citizens who lack sufficient Media and Information Literacy (MIL) competencies to distinguish them. Satire and parody can, in fact, serve as effective counters to disinformation by highlighting the absurd elements of disinformation (and those who create and disseminate it) in effective and engaging ways. However, these communications practices should not generally be treated as constituting disinformation.

[10] A post designed to provoke emotional response in its readers (e.g. anger, compassion, sadness, fear), and thus causes the user to stimulate further engagement (i.e. 'click') by following the link to the webpage, which in turn generates ad views and revenues for the website owner. The defining characteristic of clickbait is that it fails to deliver on the headline, meaning the 'clicker' has taken the bait but the article will not fulfil expectations.

[11] A form of psychological manipulation: https://dictionary.cambridge.org/dictionary/english/gaslighting

1.2 Definitions and scope

There are many different and somewhat contradictory definitions of disinformation, and whether and how it relates to misinformation. The conceptualisations generally share the trait of falsity as an essential criterion, with the result that the terms mis- and dis-information are often used synonymously and interchangeably (e.g. in Alaphilippe, Bontcheva et al., 2018b).

For its part, the Oxford English Dictionary[12] (OED) appears to distinguish the labels on the basis of one being linked to the intention to deceive, and the other the intention to mislead (although it is not clear how these objectives differ):

- **Misinformation**: False or inaccurate information, especially that which is deliberately intended to deceive.
- **Disinformation**: False information which is intended to mislead, especially propaganda issued by a government organisation to a rival power or the media.

This definition also links one of the terms (disinformation) to a particular actor (governmental), which would seem to suggest a narrowing of the scope of its remit. Others have defined disinformation specifically in the context of elections, as "*content deliberately created with the intent to disrupt electoral processes*" (Giglietto et al., 2016). This definition is likewise too narrow for the wider variety of disinformation considered in this study.

A further perspective is evident in the recommendations of a report produced by the EU High Level Expert Group on Fake News and Online Disinformation, which includes references to possible for-profit disinformation as part of what is covered by the term:

> *"Disinformation....includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or [generate] profit. (Buning et al., 2018)."*

But intention to profit is a potentially limiting factor. For example, satire is created for profit by television producers and satirical magazines, and it would be problematic to include this communications practice as disinformation per se.

The widely-adopted *information disorder* theoretical framework (Wardle, 2017a; Wardle & Derakhshan, 2017) distinguishes mis- and disinformation as follows:

- **Misinformation**: false information that is shared inadvertently, without meaning to cause harm.
- **Disinformation**: intending to cause harm, by deliberately sharing false information

The underlying criteria in this framework could be represented as such:

[12] https://www.lexico.com/definition/misinformation

	Awareness of falsity	Underlying intent
Disinformation	Aware	"Bad"
Misinformation	Unaware ("inadvertent")	"Good / neutral"

These definitions broadly align with those in the Cambridge English Dictionary[13], where disinformation is defined as having intention to deceive, whereas misinformation is more ambiguous.

Most definitions share the feature of intentionality regarding harm (implicit in the OED semantics is that both deception and attempts to mislead are negative practices).

At the same time, operationalising a distinction based on intention (and awareness of falsity) is complicated by the fact that the motivation and knowledge of the information source or amplifier may often not be easily discernible, not only by algorithms, but also by human receivers (Jack, 2017; Zubiaga et al., 2016). There is also a risk of a Manichean assumption about who is "a bad actor", which can greatly over-simplify the situation, and entail highly subjective or problematically partisan interpretations of what and whose interests are intended to be harmed.

What this highlights is the challenge of placing intentionality and awareness of falsehood at the core of the definition of what should count as disinformation, in the face of a wider phenomenon of false or misleading content. This partially explains why some writers (eg. Francois, 2019) approach the issue not by intention (or agent awareness) in the first instance but instead by putting attention on visible behaviours such as coordinated operations involving bots (which may suggest harmful intention and awareness of falsity at play). It is the case that orchestrated behaviours (including by inauthentic actors) can signal false content, yet potentially harmful falsehoods can also spread without special amplification, and they all too often originate from authentic actors like celebrities and politicians, as shown in research (e.g. Brennen et al 2020; Satariano & Tsang 2019). At the same time, truthful content may be circulated through various behaviours and actors as part of an information or counter-disinformation campaign, which is distinct from what is recognised in regard to decontextualised or falsely contextualised content in the term 'malinformation' by Wardle & Derakshan (2017). For these reasons, it would be limiting to reduce the scope of this study to treating disinformation as if the phenomenon was defined essentially by behaviours (as much as they may often be a flag for problems).

For its part, because this study seeks to cover the wide range of responses in play around the world, it avoids a narrow approach to defining disinformation. Accordingly, it uses the term disinformation generically to describe false or misleading content that can cause specific harm - irrespective of motivations, awareness, or behaviours. Such harm may be, for example, damage to democracy, health, minority and disadvantaged communities, climate challenges, and freedom of expression. Here, therefore, the operational approach to what constitutes disinformation (and hence responses to the phenomenon) are the characteristics of falsity and potentially negative impact on targets, rather than the intentionality, awareness or behaviours of its producers(s) or distributor(s). Further, if we understand misinformation in the narrow sense of inadvertent sharing without intent to cause harm, it is evident that the content at hand often owes its origin to others' deliberate acts of disinforming citizens with harmful intent. Acknowledging this 'source' of

13 https://dictionary.cambridge.org/dictionary/english/misinformation

much of the harm is a strong reason for adopting 'disinformation' as the generic term in this study, rather than 'misinformation'.

This approach is not to be reductionist in the sense of categorising all content as either potentially harmful disinformation or (true) information which does not inflict harm. Opinion reflecting values and attitudes is one example of content that cannot be classed as true or false. Science and policy, as another example, are matters in process which evolve over time and may, at least initially, resist a binary assessment. For its part, disinformation, by its nature, claims as 'true' not only falsehoods but also often what is the category of the unknown, while frequently seeking to discredit as 'false' that content that has been definitively proven to be true - such as the overwhelming scientific consensus on climate change. It is because of the existence of genuine grey areas, that there are risks in any steps taken to counter disinformation which disregard the large realm of unknowns which exist between proven truth and demonstrated falsehoods. Such measures can stifle legitimate debate and other forms of expression which are needed to help assess the veracity of particular content over time.

The use of disinformation as a generic term applied to assess responses to false content does not preclude recognition that these responses may vary according to the diverse motivations (financial, political, ideological, personal status, etc) or behaviours of the implicated disinformational instigators and actors. For example, education is a partial remedy for *misinformation* (when understood to refer to unwitting creation or circulation of falsehoods without ill intent or awareness that the content is not true), while regulation to stop money-making from scams is one of the ways to reduce the supply of *disinformation* (using the latter term here in the narrow sense to refer to conscious and deliberate lying). Deliberate distortions and deception may be more prevalent in political and electoral contexts, while misinformation (in the narrow sense) is possibly a greater factor in the case of anti-vaccination content. The underlying theory of change entailed within a given response, is thus often linked to assumptions about intent and related behaviours. Nevertheless, especially in the context of elections, referenda, and pandemics like COVID-19, the harmful impact of false content, irrespective of intentions, and irrespective of the range of behaviours underlying them, is potentially the same. People are disempowered and serious impacts can result. So, interventions need to be appropriately calibrated.

Given the remit of this study, it makes sense for the semantic framing to use the term 'disinformation' as a meta-label to cover falsehoods (encompassing misleading messages) within content and which are associated with potential societal harm (such as negative impacts on human rights, public health and sustainable development). It is this that enables the wide-ranging unpacking of the responses to disinformation underway worldwide. The intent, therefore, is not to produce yet another definition of what disinformation is, but to provide for a broad umbrella conceptualisation of the field under examination and analysis. On this broad foundation, the research that follows takes care to signal, where appropriate, how various stakeholders responding to disinformation interpret the phenomenon, implicitly or explicitly - in regard to the particular type of response under discussion.

Adopting such an approach, this study is able to show how the complex disinformation phenomenon is being met with varying responses around the world, and the bearing that these responses have on freedom of expression and sustainable development. At the same time, it is worth highlighting how this study perceives what disinformation is not. Accordingly, disinformation should not be reduced to falsity with potential harm only in news content (as is implied in the label "fake news") and, as elaborated below, it should also not be conflated with propaganda or hate speech.

1.3 Conceptualising the life-cycle of disinformation

In order to fully understand the responses that seek to counter online disinformation effectively, it is necessary to focus not only on the message itself and its veracity, but also to investigate all aspects of the disinformation lifecycle, including its spread and effects on the target recipients.

One conceptual framework is called the 'ABC' framework, distinguishing between Actors, Behaviour and Content. This attempts to give attention to 'manipulative' actors who engage knowingly in disinformation, to inauthentic and deceptive network behaviour such as in information operations, and to content that spreads falsehoods (using manipulated media formats), or that which may be factual but is inflammatory (Francois, 2019; Annenberg Public Policy Center, 2020). The motivation here is to encourage responses to avoid acting against content that may be 'odious' but which should qualify as protected speech in a democratic society. It therefore points attention to the issue of whether responses should better focus on A and B more than C.

'AMI' is another conceptual approach (Wardle & Derakhshan, 2017), which distinguishes between:

- the **Agents**, i.e. the authors or distributors of disinformation and their motivations;
- the **Message**, i.e. the false and/or manipulated content that is being spread; the way it is expressed, and the techniques used to enhance its credibility;
- the **Interpreters** (or **Targets**), i.e. those targeted by the disinformation campaign and the effects on their beliefs and actions.

In this study, these two frameworks are adapted and converged to form a new framework that also reflects two other elements which give further insight into agents, behaviours and vehicles concerning disinformation:

- The original *instigators* of disinformation, who may be different to the agents. These are the actors who initiate the creation and distribution of this content, often harnessing and paying for operationalisation. They are the real source and beneficiary of much disinformation. In some cases, the instigators can be the same as the actual implementing agents, but in many large-scale cases the latter may be paid or voluntary supporters or contractors, as well as unwitting participants. However, the functions of instigation and agency are distinct.
- The *intermediaries* that are vehicles for the message (e.g. social media sites and apps) - which allows for attention to the key role that they play in the dissemination and combating of disinformation, and how their systems may enable - or disable - implicated content, actors and behaviours.

This aggregation can be described with reference to **1. Instigators 2. Agents 3. Messages 4. Intermediaries 5. Targets/Interpreters** - creating the acronym **IAMIT**. This approach aims to capture the complete lifecycle - from instigation and creation to the means of propagation to real-life impact, through answering the following questions:

1. **Instigators:**

    - Who are the direct and indirect instigators and beneficiaries? 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.? This includes, where discernible, if there is intent to harm and intent to mislead.

2. **Agents:**

    - Who is operationalising the creation and spreading 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. **Message**:

    - What is being spread? Examples include false claims or narratives, decontextualised or fraudulently altered images and videos, deep fakes, etc. Are the 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? Do the responses recognise that "...free *speech* does not mean free *reach*" because "there is no right to algorithmic amplification" and content moderation which may include limiting amplification should not be equated with the demise of freedom of expression online (DiResta 2018)? 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).

    - Are there real-life impacts through actions? For example, such as influencing votes, promoting protests, inciting hate crimes, attacking journalists, and providing dangerous or erroneous medical advice, raising the question of whether responses engage with the wider context or are limited to the realm of the online content at hand.

Using this hybrid '**IAMIT**' framework as a starting point for conceptualising disinformation, it is then possible to categorise responses to disinformation on this basis. In particular, we can distinguish:

- Responses aimed at the instigators and agents of disinformation campaigns (Chapters 5.1, 5.2, and 5.3).

- Responses aimed at identifying disinformation, i.e. verifying messages in terms of falsity, exposing the instigators and agents. (Chapters 4.1, 4.2)

- Responses aimed at curtailing the production and distribution of disinformation and related behaviours, implemented particularly by intermediaries (Chapters 6.1, 6.2, and 6.3).

- Responses aimed at supporting the targets/interpreters of disinformation campaigns (Chapters 7.1, 7.2, and 7.3).

1.3.1 Disinformation and propaganda

Disinformation, as unpacked above, has distinctions from, and overlaps with, the notion of propaganda. Intentionality is core to an understanding of propaganda, in that the latter implies an organised, orchestrated campaign. This is not always the case with disinformation as broadly conceptualised in this study.

At the same time, as noted in the OED definition above, and in the Joint Declaration on Freedom of Expression and 'fake news', disinformation, and propaganda[14], disinformation may overlap with propaganda, which:

[14] https://www.osce.org/fom/302796

> *...is neutrally defined as a systematic form of purposeful persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for ideological, political or commercial purposes through the controlled transmission of one-sided messages (which may or may not be factual) via mass and direct media channels. (Nelson, 1996: p232-233)*

There is a long history where propaganda and disinformation are intertwined (Posetti & Matthews, 2018). Techniques of deceitful or 'dark' propaganda (e.g., selective use of facts, unfair persuasion, appeal to fear) are employed widely, e.g. in anti-EU campaigns, post-truth politics (Keane, 2018), ideology-driven websites (e.g., misogynistic or Islamophobic), and hyperpartisan media. This is often with the intent to effect actual behavioural changes e.g. to deepen social division, increase polarisation, influence public opinion, or shape key political outcomes.

While propaganda typically puts the emphasis on strategic persuasion towards action through harnessing narrative, identity and emotionality, the intellectual 'work' of disinformation (as conceptualised here) is to 'mess' with facts and knowledge in the primary instance (rather than target attitudes or behaviours). Propaganda typically harnesses disinformation to reinforce its bigger purpose. Yet, while disinformation can make a significant contribution to a propaganda package, these are not only analytically distinctive interventions, each can also stand alone. What complicates assessment is when disinformation is fused with propaganda techniques around linguistic, cultural, and national differences, such as to create new social barriers and amplify divisions. This fusion technique is a notable feature of divisive political campaigns, whether conducted internally within a State (e.g. campaigns with nationalistic objectives), or by foreign actors (e.g. designed to destablise political life in another State).

The rationale behind combining propaganda techniques with disinformation campaigns is to *enhance the credibility* of the message. It must be emphasised that the credibility of a message is separate from its veracity, since the former is about subjective perception of whether specific information seems credible, whereas verification is about evidence-based, independent assessment.

In addition, the merging of propaganda techniques and disinformation can be a strategy to move away from the use of patently false content in favour of using decontextualised, manipulative, and misleading content in order to distort the information ecosystem.

1.3.2 Disinformation and hate speech

Hate speech relies on group 'othering', and may engage disinformation as part of its arsenal, such as by reinforcing its message with false information and generalisations about a particular class of persons. Such speech may be part of a propaganda initiative, although not necessarily.

An important distinction needs to be made between disinformation on one hand and hate speech on the other, where hate speech is understood as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (UN, 2019; see also UNESCO, 2016). The two phenomena often intersect for instance when online abuse and disinformation are used hand-in-hand, such as in political smear campaigns, or misogynistic attacks on women journalists. They are nevertheless conceptually distinct, since false information can stand by itself and not touch on hatred, e.g. in anti-vaccination disinformation. Hatred, for its part, does not necessarily always implicate disinformation: it can rely simply on expressions of opinion and incitement without falsehoods, such as by amplifying fears, xenophobia, misogyny or other prejudices.

The focus of this study, in particular, is on the range of responses to disinformation assessed through a freedom of expression lens. Therefore, responses purely focused on hate speech are out of scope. Where responses to disinformation are tied up with hate speech, however, the phenomenon is examined from that perspective.

1.4 Disinformation, freedom of expression, and the UN sustainable development goals

Seventeen global Sustainable Development Goals (SDGs) were set by the United Nations General Assembly in 2015[15]. A number of them are impacted by use of broadband technology and internet communications company services for the proliferation of online disinformation - but also for the implementation of some categories of disinformation responses. These are:

- **SDG 16 on peaceful and inclusive societies and SDG 5 on gender equality**:
  - **Online disinformation is** often used to target individuals (such as politicians, journalists, human rights defenders), governments, groups such as ethnic minorities, women and gender identity-based communities, and religious congregations and identities, including in messages which may lead to violence, hatred, and discrimination.
  - The **algorithms used by social media and search engines** to prioritise and recommend content (including disinformation) have been shown to prioritise and recommend content that is attention- and engagement- focused, and prone to bias, accordingly potentially working against inclusivity. (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018a)
- Of particular relevance to this report on Freedom of Expression (FoE) and disinformation is **SDG 16.10 on public access to information and fundamental freedoms**
  - Citizens' rights to express themselves freely and participate on an informed basis in online societal debates are jeopardised by online disinformation, especially when distributed at scale. False content can undermine citizens' beliefs and trust in facts, science and rationality, and therefore stoke cynicism about online information that contradicts their opinions. This can deter public participation, and impact negatively on the exercise of rights and obligations concerning civic actions. This is especially relevant for citizens and communities targeted in disinformation campaigns using hate speech as a tool to fuel division and inflame tensions.
  - It is noteworthy that politicians and governments are among the main instigators and vectors of disinformation (Brennen et al 2020; Bradshaw & Howard 2019).

[15] https://sustainabledevelopment.un.org/?menu=1300

- The rising use of AI algorithms for automatic content filtering of disinformation (and other kinds of content) can lead to over-censorship of legitimate content, thus infringing on the author's freedom of expression and right to access information. These algorithms can also exhibit inherent biases and be prone to manipulation.

- Orchestrated and organic disinformation campaigns targeting journalists (particularly women journalists) and news outlets as a means of undermining citizens' trust in journalists and journalism as credible and independent sources of information.

- Another example is disproportionate legal responses to disinformation which can sometimes lead to internet shutdowns and censorship, inhibiting reporting, and criminalising journalism, as well as vague legal definitions of disinformation which can be used to silence political opposition or dissenting voices (e.g. via 'fake news' laws)

- **SDG 4 on inclusive and equitable quality education:**

  - As citizens are increasingly using the internet and search engines to find information for educational purposes, **high levels of online disinformation** can seriously impact on the knowledge processes essential for the quality of education, as many learners are unable to judge the trustworthiness of online sources and the veracity of the information they find online. This has become increasingly important as COVID-19 has forced so much education online.

  - The search engine algorithms used by citizens to find information can be gamed to prioritise viral disinformation, which in turn can lead to learners (especially children and older generations) starting to believe in conspiracy theories and other false or low-quality online information.

  - On the positive side are investigative journalism projects focused on disinformation and **media and Information literacy initiatives**, **including data literacy**, designed in response to online disinformation that aim to impact positively on citizen education, knowledge, and abilities to identify and protect themselves from disinformation.

- **SDG 3 on healthy lives and promotion of well-being for all ages**:

  - Health-related disinformation in general - as demonstrated during the COVID-19 pandemic and including long-standing anti-vaccine propaganda - jeopardises citizens' healthy lives and well-being (e.g. diet-related disinformation). As a result of anti-vaccine disinformation, vaccine take-up rates have shown a sharp decline in recent years (e.g., in Africa (France 24, 2020), Asia (Power, 2020), Europe (Larson, 2018) and North America (Burki, 2019)).

Disinformation runs counter to the agreed SDGs. Yet, its purveyors (wittingly or unwittingly) and operating with a range of motives, still foresee advantage in investing time and resources in producing and circulating it - leveraging the business models and technologies of internet communications companies and the news media to do so.

> *"At the same time, disinformation is a 'game' with no long-term winners. Escalating the volume of disinformation in play ultimately devalues facts for everyone and puts humanity on a path towards ubiquitous ignorance. The achievements of civilisation based upon freedom of expression to date are being jeopardised as a result. At stake are issues of health, democracy, financial security, the environment, peaceful resolution of social conflict, social cohesion, and more."*

Disinformation is a phenomenon that is too challenging for any single state or company to manage in isolation - it requires collaboration with researchers, civil society and the news media. Paid microtargeting of individuals with disinformation content is one example that calls out for unprecedented cooperation; the peer-to-peer use of closed social messaging networks that spread falsehoods is another.

It is for this reason that this study examines the range of responses that can prevent, inhibit and counter disinformation. The following chapters assess the full suite of possibilities, and their respective strengths and weaknesses, as well as potential risks to freedom of expression rights, as multiple actors seek to tackle disinformation.

The next chapter - Chapter two - introduces the typology of disinformation responses which forms the backbone of this study. Chapter three provides a detailed mapping of existing research, highlighting knowledge gaps and opportunities for further study. Then, each of the eleven response types presented in the original taxonomy devised for this study is systematically analysed.

1.5 Methodology

The findings presented here are the result of desk research carried out (September 2019-July 2020) by a multidisciplinary team of international authors who worked in a highly collaborative fashion.

The research for this study sought to include sources pertaining to countries on all continents, including where possible (according to the language capabilities of the researchers), materials in languages other than English. The libraries and databases of academic institutions, individual States, civil society organisations and news media websites were targeted by the researchers. Many of these collected sources have now been aggregated into the study's bibliography.

An Expert Oversight Group comprised of Associate Professor Fabrício Benevenuto, Federal University of Minas Gerais; Prof Divina Frau-Meigs, Université Sorbonne Nouvelle - Paris 3; Prof Cherian George, Hong Kong Baptist University; Dr Claire Wardle, Executive Chair of First Draft; and Prof Herman Wasserman, University of Cape Town provided feedback. The research team also worked closely with the UNESCO secretariat to shape this study.