Research Context and Gaps

3

Authors: Diana Maynard, Julie Posetti, Kalina Bontcheva and Denis Teyssou

Chapter 3 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 situates the disinformation focus of this report within the context of existing theoretical frameworks and prior reports on this topic. It also relates disinformation to freedom of expression and the Sustainable Development Goals relevant to the Broadband Commission (specifically SDG 16 on peaceful and inclusive societies, and SDG 16.10 on public access to information and fundamental freedoms). In particular, the focus is not only on the false content itself, but also the actors, their motivations for sharing disinformation, and the targets of disinformation campaigns, thereby including discussion of the amplification and manipulation of this kind of content. Additionally, the chapter examines the literature regarding modalities of response to disinformation. Then, it discusses in more depth the gaps in the research carried out prior to early 2020, especially in relation to defining the novel contributions of this study, compared with previous reports on the manifestations of disinformation.

With this gap analysis, special attention is paid to the impact of disinformation on societies and its reception by the public, by reviewing literature in cognitive science and social psychology, in addition to that found in the fields of politics, journalism, information and communication sciences, and law. The review encompasses not only academic literature, but also policy reports from industry and civil society groups, white papers, books aimed at the mainstream public, and online news and magazine articles. It should be emphasised, however, that this review is not intended to be exhaustive, rather it is designed to map some of the key research trends and gaps, while also identifying gaps in responses.

The chapter does not attempt to definitively assess the quality of the selected works, but rather to understand the nature of a range of existing research on disinformation, the theoretical underpinnings, types of studies that have been carried out, and the ways in which disinformation has been discussed in journalistic, academic and official (i.e. State and non-State actors) circles, as well as how this has been disseminated more widely to the general public. It then summarises the findings, focusing on some key areas such as political disinformation and policymaking, and highlights some emerging trends, before discussing the limitations of such research and situating this report within the scholarship.

The aim of this review is thus to understand what can be learnt about what kinds of disinformation exist; who instigates and disseminates it and why; at whom the information is targeted; the relative effectiveness of different kinds of disinformation and different modalities; existing responses to it; and what recommendations have been made for the future, in particular in the light of freedom of expression concerns. This paves the way for the following chapters which investigate in more depth the various responses to disinformation, as well as the theories of change associated with them, and possible benefits and drawbacks.

Attention is given to highlighting new threats, such as undermining freedom of expression by indiscriminately using Artificial Intelligence (AI) filtering methods, and to the rise of synthetic media (also called 'deepfakes') as new modes of disinformation. The latter problem can already be seen in practice, where several politicians and journalists have been targeted and smeared with inappropriate sexual misconduct allegations in manipulated and/or deepfake videos.[16]

A related recent trend, which has been largely underestimated in the past, is the rise of adversarial narratives (Decker, 2019), whereby disinformation strategies include not only simple conspiracy theories and outright lies, but also involve more complex phenomena

[16] See examples from Finland (Aro, 2016), India (Ayyub, 2015), and South Africa (Haffajee & Davies, 2017), among others.

whereby true and false information is emotionally charged and deliberately entangled in intricate webs designed specifically to confuse, shock, divert and disorientate people, keeping truth-seekers always on the defensive. If information is a condition for public empowerment, then disinformation can be seen to function in terms of displacing and discrediting information, often with the rationale of disempowerment and driving confusion. One example of this is 'gaslighting', a powerful strategy aimed at control through power and manipulation of people's perceptions of reality - thereby generating fears and sowing disruption, and then appearing to offer solutions (Keane, 2018). These disinformation techniques, often described as the "weaponisation of information", can destroy social cohesion and threaten democracy (Hansen, 2017). They can stimulate public demand for stronger certainty and greater political control, thereby risking further curbs on freedom of expression, and strengthening social authoritarianism (Flore et al., 2019). On the other hand, responses to disinformation are developing in sophistication and incorporating human rights standards in order to counter the potential harms at stake.

3.1 Conceptual frameworks for understanding contemporary disinformation

In recent years, there has been a flurry of research investigating not only the nature and extent of disinformation, but also the psychological underpinnings and theoretical frameworks. These frameworks capture different aspects. An overarching view is taken by Wardle and Derakhshan (2017), who consider 'information disorder' as a tripartite problem where (in their definitions) 'disinformation' sits alongside 'misinformation' and 'mal-information'. Other views range through to narrower classification systems such as the political disinformation campaign characterisation of the Digital Forensic Research Lab (Brooking et al., 2020).

3.1.1 "Information disorder" and "information warfare"

In their report for the Council of Europe, Wardle and Derakhshan (2017) elaborate their concepts and provide a background summary of related research, reports and practical initiatives produced around this topic up to the middle of 2017. Their report investigates ideas and solutions for, and from, the news media and social media platforms, as well as examining future directions and implications. This includes focus on the use of AI, not only for detecting disinformation but also for creating it. The report also details 34 recommendations for technology companies, governments, media organisations, funding bodies, and broad citizenry. Many of these recommendations are already in place in some form (for example, some technology companies are already building fact-checking tools. Some recommendations lend themselves to further unpacking (for example, how civil society could "act as honest brokers", or how education ministries could "work with libraries").

Wardle and Derakhshan's conceptual framework follows on from their previous work "Fake News, It's Complicated" (Wardle, 2017a), which defines seven types of mis- and dis-information, ranging from satire and parody (which, being mis-interpretable, have the potential to lead to what they call mis-information) through to full-blown deliberate fabrication. The framework situates the production and distribution of disinformation

as a tripartite process consisting of Agent, Message, and Interpreter (target). However, as signalled in the Introduction to this study, the practicality of this frame encounters the challenge of distinctions between mis-information and dis-information being based primarily on motive and awareness of falsity. Motives are not only diverse and often contradictory, but also frequently not clear. Furthermore, the distinction may risk over-emphasising intentionality at the expense of commonality of effect. For example, if people decide against vaccination through engagement with false content, the consequence is the same, whether the mode of transmission is mis-information or dis-information. Where motives become significant as an issue, although they are often hard to pinpoint, is in assessing the appropriateness of a given response with respect to how it establishes the issue of motives at hand. That is why this study pays attention to investigative responses as a source of knowledge for informing other types of responses.

Another consideration related to the 'information disorder' framework is that it can favour a binary distinction between information that is 'ordered' or 'disordered', and thereby reinforce a view that reduces the veracity of communications to issues of black and white, while overlooking or denying the range of unknowns, whether these be scientific uncertainties or legitimate policy debates. Another issue is that 'mal-information' could be interpreted in a way that stigmatises a range of non-disinformational narratives, which intrinsically select and interpret particular facts and contexts as part of their legitimate contestation around meaning.

In this light, the research in this study operates at a more abstract level than privileging categories of false or misleading content through the criteria of motives, and instead puts the focus on all false content that has the potentiality of defined harm. This provides a means towards assessing the full range of responses as they conceptualise themselves.

A strategically focused approach to the issue of disinformation is assessed by Derakhshan (2019) in his report "Disinfo Wars". This discusses the relationship between agents and targets in what he calls a "taxonomy of information warfare". Accordingly, the approach directs the idea of disinformation into a much narrower concept that articulates to political and even military strategy. An example of the latter is the perspective on 'Information Operations' / 'Influence Operations' taken by the Rand corporation, which links these terms to "the collection of tactical information about an adversary as well as the dissemination of propaganda in pursuit of a competitive advantage over an opponent".[17] A similar position is adopted by the European External Access Service (EEAS) East Stratcom Task Force[18]. Derakhshan argues that the majority of money and effort spent on countering disinformation in "information warfare" should be focused on those who are targeted, i.e. non-state actors like the media.

While his argument covers a wide range of activities, it focuses to some extent on false content distributed with a particular motive, as with Wardle's earlier work (Wardle, 2017a). As discussed above, this is complicated operationally, and it goes beyond even the complex issues of attribution. In addition, while strategic focus on geopolitical dimensions and particularities is important, society also faces the issue of disinformation as a far wider problem. There is also a lack of evidence that work with one constituency (the media, or the general public) is less or more effective than work with another.

17 https://www.rand.org/topics/information-operations.html
18 https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1006

3.1.2 Political disinformation campaigns

A perspective that relies less on warfare metaphors but deals with political disinformation as a broader concept has been adopted by researchers at the Digital Forensic Research Lab and Google's Jigsaw (a division that includes a focus on combatting the 'unintended consequences' of digital technology) has proposed and tested a classification system for political disinformation campaigns, built on 150 variable options (Brooking et al., 2020). The main aim of this framework is to enable the description and comparison of very different kinds of political disinformation efforts. The scheme has six major categories: target, platform, content, method, attribution, and intent. Each of these is broken down into further categories and subcategories. The table below shows the first and second level categories, with some examples of the third level. Typically, the third level categories are binary (e.g. whether it is a government-related target or not), although the quantitative measures involve numbers or ratios, and some have free-form responses. In addition, all second level categories have a category where free-form notes can be added, and some also have an "other" subcategory.

1st level	2nd level	Notes / Examples
Target	Primary target	Government, political party, business, racial group, influential individuals (including journalists) and groups of individuals, etc.
	Quantitative measures	Indicators/rankings of political stability, internet freedom, refugee counts etc.
	Concurrent events	War, elections etc.
	Secondary target	Rarely used
	Tertiary target	Rarely used
Platforms	Open web	State media, independent media, other
	Social media	Facebook, Instagram, Twitter, forums, etc.
	Messaging platforms	WhatsApp, Telegram, Wechat, SMS, etc.
	Advertisements	(Purchased by disinformants to disseminate a message of disinformation, including on social media and the open web)
	Email	
Content	Language	
	Topics	What the disinformation is about, e.g. government, military, elections, terrorism, racial, etc.
Methods	Tactics	Brigading, sock puppets, botnets, search engine manipulation, hacking, deepfakes, etc.
	Narrative techniques	Constructive (e.g. bandwagon, astroturfing); Destructive (e.g. intimidation, libel); Oblique (trolling, flooding)
Attribution	Primary Disinformant	Country, bloc, other
	Disinformant Category	As for target category, e.g. government, political party, business, influential individual, minority group
	Quantitative Measures	As for target, e.g. political stability data, internet freedom, refugee counts
	Concurrent Events	As for targets, e.g. war, elections, etc.
	Secondary Disinformant	Rarely used
	Tertiary Disinformant	Rarely used

Intent	Object	Free text (1 or 2 short sentences)
	Category	e.g. civil, social, economic military

Table 1. *Simplified version of the political disinformation campaign categorisation scheme devised by Brooking et al. (2020).*

In this work, Brooking et al. define political disinformation as "disinformation with a political or politically adjacent end", which captures "disinformation spread in the course of an election, protest, or military operation, as well as "the widespread phenomenon of political 'clickbait' disseminated for personal financial gain".

Their framework defines a political disinformation campaign as "political disinformation that demonstrates both coordination and a discrete objective." They note that, first, objectives may not always be obvious, even though they must exist; and second, that campaigns with changing objectives can thus become discernibly distinct from each other (i.e. if the objective changes, it becomes a new campaign). Furthermore, they note that political disinformation campaigns almost always involve what they call "amplification of content". This concept, which is discussed in more detail in the following section, is termed "political astroturfing" by Keller et al. (2019), "coordinated inauthentic behavior" by Facebook (Gleicher, 2018a), and noted as a feature of 'astroturfing' in the targeting of journalists with misleading information designed to "mislead, misinform, befuddle, or endanger journalists" by Posetti (2013). Not all instances of this constitute disinformation as such, but there is a clear overlap since the aim is to create an "illusion of consensus or popularity," and in some instances, to inflict harm. Some researchers have tried to capture this complex interplay through a "matrix of disinformation harms", which encompasses polarisation and radicalisation along one dimension and propaganda and advertising along the other (Frau-Meigs, in press).

In providing a basis for comparing different kinds of disinformation, this framework also has the benefit of enabling detailed background information to be represented. Understanding the situational context such as the presence of military conflict, or levels of political stability may help with both short and long-term assessment and the provision of appropriate solutions. However, it also risks the case that some of the factors may be unknown or irrelevant. As with other frameworks discussed, notions of intentionality and attribution are also not always evident. As significant as deliberate disinformation is during such political campaigns, this study bears in mind the wider picture that includes unintentional falsehoods in play (such as health issues), and therefore maintains a focus that covers responses wider than those dealing with political issues.

3.1.3 Information influence

Similar to the political disinformation campaign characterisation, the Handbook for Communicators (Pamment et al., 2018) views disinformation in the context of the wider sphere of "influence activities" and from the point of view of policymaking (in the case of that handbook, the Swedish government). This framework deconstructs influence activities conducted by foreign powers, focusing on rhetorical strategies, techniques, and influence stratagems, and aims to enable policymakers to identify, understand, and counter these increasingly sophisticated activities and campaigns. This approach focuses particularly on safeguarding society's "democratic dialogue", which they explain as "the right to open debate, the right to arrive at one's own opinion freely, and the right to free expression". In this light, they view methods of social resilience, such as informing and educating the public, as the foundation for combatting disinformation and influence

activities, and they focus their attention on public communicators within governments and state organisations accordingly.

'Information influence' in this framework is closely related to disinformation, which Pamment et al. define as "a technique based on the distribution of false information intended to mislead and deceive". The authors argue that those who conduct "influence activities" are only a step away from (perfectly legitimate) advertising campaigns which attempt to sway people to buy a product, for example. They argue that it is precisely the notion of openness that differentiates them: advertising and public relations should be transparent in their motives, and follow clear rules; on the other hand, information influence involves the covert and deceptive deployment of false content. In this regard, the approach of Pamment et al. overlaps substantially with broader uses of the term 'information operations' such as as references to the combination of co-ordinated and inauthentic behaviour (such fake profiles and hidden behaviours) as a wider phenomenon than cases of military or geopolitical deployments.

> *"Given that societies are built on trust, deceptive 'information influence' undermines the democratic principle of the public's right to know and access information"*

Given that societies are built on trust, deceptive 'information influence' undermines the democratic principle of the public's right to know and access information, potentially destabilising democracy by muddying the informational waters so much that it becomes impossible to discern accurate information from falsehoods, and credible journalism from propaganda, broadly undermining trust in public interest information. In this regard, the concept of 'information influence' also resonates in part with the concept of infodemic[19] popularised by the World Health Organisation, and which designates "an overabundance of information – some accurate and some not – occurring during an epidemic. It makes it hard for people to find trustworthy sources and reliable guidance when they need it."[20]

The theory of information influence adopted by Pamment et al. has three parts: awareness, identification, and countering.

Awareness consists of understanding the anatomy of an information campaign, as well as the process of opinion formation. In this light, information influence can be distinguished by three main features: it is deceptive, intentional, and disruptive. It should be noted, however, that these aspects are not always easy - or even possible - to determine, signalling an important gap in this theory. As previously discussed, intentionality can be hard to determine, or at least to attribute, and the extent and impact of disruption is hard to measure.

The process of **identification** of 'information influence' is based on the idea of strategic narratives, which can be seen as a deliberate manipulation of some fundamental belief such as that the earth is round.[21] Distinct from other frameworks such as those of Derakhshan (2019) and Brooking et al. (2020), target groups here are always the public,

[19] https://www.merriam-webster.com/words-at-play/words-were-watching-infodemic-meaning; https://www.who.int/teams/risk-communication/infodemic-management

[20] https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference

[21] Note that since disinformation, as conceptualised in the approach of this study, could count as one of a number of different types of information influence, this does not signify that all strategic narratives equate to disinformation, nor that all strategic narratives are fundamentally deceptive.

and can be broken down into general public, socioethnic groups (e.g. a religious group), and psychographic groups (those with specific personality traits).

In the framework of Pamment et al., disinformation is defined far more narrowly than it is treated in this report. It is classified as a technique distinct from techniques involving technical exploitation, which includes bots, 'deepfakes', and sock puppet networks. These in turn are seen as distinct from the category of "deceptive identities", which includes what they term "fake media"[22] and the loosely defined "counterfeits". The other three categories - social and cognitive hacking, malicious rhetoric, and symbolic actions, are more loosely related to disinformation, encompassing notions of bias such as filter bubbles, strawman tactics, and leaking and hacking, respectively. On the other hand, satire and parody are (problematically) classified as disinformation. In contrast, in this study, it is recognised that while disinformation is often orchestrated, it is not per se a technique – instead, it makes use of techniques like technology and deceptive identity. The same point applies to the analysis of Francois (2019), which comes close to elevating behaviours, including inauthentic behaviours (and fake actors), to being defining features of what should be considered as disinformation. While such trademark signs of disinformation are significant, this study also recognises that many cases of disinformation also exist without these features.

The framework by Pamment et al. faces the challenge, like many already discussed, of the practical ability to make distinctions given reliance on assumptions about motive and intent. This challenge also applies to those who interpret behaviours as a barometer of motives, in that there are complex levels between, for instance, a person who shares false content believing it to be true and helpful, and an agent who amplifies it, and further compared with an instigator operating with a wider strategy. On the other hand, the Pamment et al. assessment does avoid a potential pitfall of the concept of 'mal-information', in recognising that not all persuasive or strategic narratives equate to disinformation.

Finally, in terms of strategies for **countering** information influence, Pamment et al. suggest four categories, ordered temporally. The first responses are the two fact-based techniques of assessment and then informing. These are followed by two advocacy-based techniques and, lastly, defence. The first step, assessing the situation, can involve methods such as fact-checking and investigating the transparency of the information. Informing involves steps such as making statements to signal issues, and correcting factual inaccuracies. Advocating is described as use of mechanisms such as dialogue and facilitation. Defence is the final stage in the process which involves official blocking, reporting, and removal of disinformation. While not approaching the extent of responses covered in this study, the Pamment et al. framework does have the merit of highlighting the links between awareness, identification and response.

[22] Editors' note: The terms 'fake news' and 'fake media' are problematic and should be avoided where possible because they are frequently weaponised as tools in the disinformation 'arsenal' in an attempt to discredit journalists and news reporting by actors seeking to chill accountability journalism. See UNESCO's *Journalism, 'Fake News' and Disinformation* for further discussion (Free to download here in multiple languages: https://en.unesco.org/fightfakenews)

3.2 Empirical and applied research

Moving on from theoretical frameworks which attempt to define and classify various kinds of disinformation and, in some cases, potential responses to it, this chapter now focuses on more empirical and applied research, looking at some key trends and examples of specific case studies.

Bradshaw and Howard's "Global Inventory of Organised Social Media Manipulation" (2019) focuses on social media manipulation by governments and political parties. Their report analyses the trends of what they call 'computational propaganda', looking at tools, capacities, strategies, and resources. Their surveys show that in recent years, evidence of organised social media manipulation campaigns is becoming more widespread worldwide, with the number of countries involved increasing by 150% in two years. In 2019 they found evidence of such campaigns in 70 countries, up from 48 countries in 2018 and 28 countries in 2017, with Facebook being the most common social media source.

Martin and Shapiro (2019) also present a detailed classification system for online "foreign influence" initiatives, which compares the characteristics, tactics, rhetoric and platform choices of different attackers. A few studies have attempted to dig deeper into the underlying motives of these kinds of initiatives, but these are restricted to country-specific case studies. Ong and Cabañes (2018) investigate, from an "ethnologically informed" perspective, the motivations and behaviour of those who are recruited to produce networked disinformation and social media manipulation in the Philippines, while Chaturvedi (2016) investigates similar issues in India.

However, despite these and other reports discussing these forms of organised political disinformation and 'influence operations', there remains a lack of coordinated in-depth research into this phenomenon as a whole, especially at more than a case- or country-specific level. These systems can influence people sufficiently to change their votes, buy products and change perceptions - sometimes with enormous consequences for democracy or public health. So-called 'dark PR' has been defined as the "manipulation at scale for money without any concerns for the damage to the planet, country, or even individual safety"[23], leading to a worldwide industry of PR and marketing organisations buying services that use fake accounts and false narratives to spread disinformation via end-to-end online manipulation systems (Silverman et al., 2020).

A number of countries around the world have sought to make it a crime to create and distribute disinformation of this type (Adhikari, 2020), although the definitions of what is acceptable vary substantially. In practice, finding the sources and proving intent may not be a trivial process for either law enforcement agencies or companies themselves. Adhikari notes that Facebook has attempted to curb such disinformation spreading practices, banning in 2019 a number of dark PR firms for attempting to influence elections, or for what it calls "coordinated inauthentic behavior" in various countries. However, these kinds of activity are still widespread, and new companies promoting such services can be easily set up.

[23] Definition by Rob Enderle, principal at the Enderle Group, quoted in the E-commerce Times article '*Black PR' Firms Line Their Pockets by Spreading Misinformation* by Richard Adhikari: https://www.ecommercetimes.com/story/86444.html

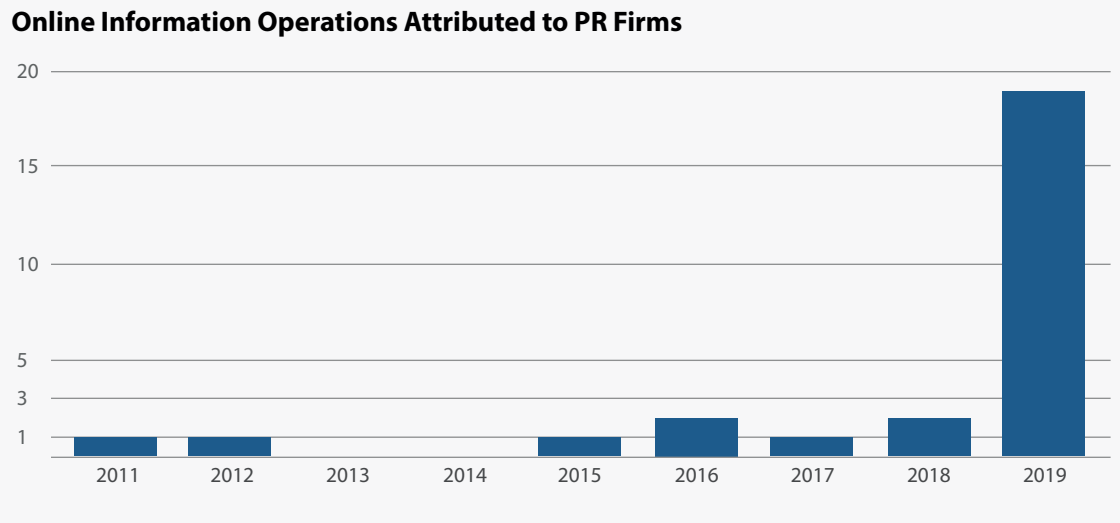

Figure 3. *Chart Source*: *Buzzfeed News* (*Silverman et al., 2020*)

From the BBC, analysis of misleading stories during the COVID-19 pandemic resulted in a typology of seven kinds of actors involved in generating disinformation (Spring, 2020). Other noteworthy journalistic investigations giving insight into the agents and instigators include those from Buzzfeed, Codastory and Rappler, for example (Dorroh 2020) - as discussed further in chapter 7.1. At the time of writing, however, there was a scarcity of detailed academic studies on this phenomenon, and methods for preventing it at its source were not obvious.

While notions of 'influence operations' are not themselves new, the proliferation of these in 2019, as illustrated below, requires urgent attention. On the other hand, it should be noted that not all 'influence operations' necessarily equate to the characterisation of disinformation used by this study, in the sense that some such initiatives may not harness false or misleading content, nor rely on inauthentic behaviour. Recent high profile cases concern mechanisms such as "inciting racial tension" (Neate, 2017) and "co-ordinated inauthentic behaviour" (Gleicher, 2019a) which leave open a number of possibilities as to their harnessing of disinformational content. Some coordinated campaigns can be mounted with accurate content, transparent behaviours and authenticated actors, as for example in advocacy by some civil society groups, public health communications by governments, and public relations initiatives by companies. The topic of organised influence therefore needs to be approached with appropriate nuance when researched from the point of view of when and how it intersects with false and misleading content with potential to harm.

3.2.1 Social and psychological underpinnings

A strand of research into disinformation situates it within its social and psychological context in order to define and understand appropriate responses. Even if some of the mechanisms of disinformation are new, responses to them can/may be guided by the decades of research into human cognition. As will be discussed elsewhere in this study, it can be hard to persuade people who want to believe a piece of information that this content is indeed false - or that a false 'fact' can make a difference to the meaning they attribute to a bigger picture. Even if fact checkers disprove false information, research has shown that it can be extremely difficult to change people's minds on misconceptions, especially if they believe there is even a kernel of truth within the falsity (Johnson & Seifert,

1994; Nyhan & Reifler, 2010). As the economist J.K. Galbraith once wrote: "Faced with a choice between changing one's mind and proving there is no need to do so, almost everyone gets busy with the proof" (Galbraith, 1971). Repetition and rhetoric are powerful devices: people are more likely to believe information when they see it repeated over and over again (Zacharia, 2019). Importantly, according to Effron & Raj (2019), such repeated exposure means that such people also have fewer ethical concerns about resharing it, whether they believe it or not.

Longstanding research in political science has shown the power of rhetoric time and again (Kroes, 2012; Grose and Husser, 2008): linguistically sophisticated campaign speeches by election candidates are far more likely to influence people to vote for them. This linguistic sophistication involves presenting the message - no matter what its content - in a tailored rhetorical way that also conveys emotional resonance. However, one finding has been that while linguistic sophistication (i.e. presenting the message in a particular rhetorical way, rather than changing the message itself) is more likely to persuade those with higher education, it does not dissuade those without (Grose and Husser, 2008). While campaign speeches as such should not be equated with disinformation, these findings lead to the observation that disinformation combined with non-informational dimensions (emotional quotients) could be more powerful than when it is presented alone.

Taking this a step further, others frame relevant aspects of disinformation within the notion of "psychological warfare" (Szunyogh, 1955, cited in Cordey, 2019). Guadagno and Guttieri (2019) provide an overview of research in psychology and political science in this context through the spreading of disinformation. They review a number of social, contextual and individual factors that contribute to its proliferation. Focusing specifically on the spread and influence of 'dark propaganda' online, they consider the social elements such as online interactions, and the technological affordances that affect this. They also situate disinformation in the context of other media-related factors that might contribute to or drive the spread and influence of disinformation. However, their research focuses only on two specific case studies, (the United States[24] and Estonia). While they find differences between these cases, their research findings cannot necessarily be extrapolated to a wider geographical or situational sphere. Alongside these notions, it is useful to understand some of the reasons why people believe false content, and why they share it even when they know or suspect it is not true. A number of studies have been conducted concerning the psychology of belief, leading to the argument that behavioural sciences should play a key role in informing responses to disinformation (Lorenz-Spreen et al., 2020). Lewandowski looks specifically at conspiracy theories such as those around the coronavirus, claiming that in a crisis, people typically go through various stages of denial including not believing there is a crisis, blaming others for it, or not believing solutions will work, all typically leading to the support of these conspiracy theories (Cook et al., 2020).

In countries whose mainstream media is largely or fully controlled by government authorities, there is often a public distrust of such sources, particularly where this is linked with historical or current issues such as apartheid and corruption. In such countries, "radio trottoir" (literally, pavement radio) (Ellis, 1989) and other forms of underground media are often seen by the public as more trustworthy than official sources of information (Wasserman, 2020). Wasserman's study conducted in sub-Saharan Africa (Wasserman & Madrid-Morales, 2018) found low levels of trust in the media, a high degree of exposure to misinformation, and that people often contributed to its spread even with the knowledge

[24] The U.S. withdrew its membership from UNESCO in October 2017.

that facts were incorrect, to a much greater degree than United States (U.S.) citizens. This finding highlights the need to assess the extent to which strategies to counter disinformation should go beyond basic educational and media literacy strategies in order to tackle the root causes of mistrust.

The work of Wasserman and Ellis, among others, indicates that the reasons for knowingly sharing false information are likely to be connected with the notion of group allegiance. In other words, notions of truth are less important than notions of solidarity, and as long as a piece of information aligns with our world view, we often do not investigate its factuality. A study by Pennycook & Rand (2019) found distinct differences between people's ability to distinguish true from false information and their likelihood of sharing that information - in other words, it was not only the information they believed to be true that they said they would share. It is clear from all these findings that not only do "cognitive miserliness"[25] and cognitive bias play a part in our believing and sharing of false information, especially in an information-rich environment, but also that we are driven by heuristics such as social endorsement, and these elements should therefore be a factor in assessing responses to disinformation.

In order to respond effectively to disinformation, it is also important to understand some of the reasons why people are reluctant to change their opinions even when faced with evidence to the contrary. Hans Rosling discusses the notion that people typically have a number of negative misconceptions about the world (such as life expectancy in poorer countries, or the death rate from natural disasters), and even when faced with figures that disprove these, people struggle to accept them (Rosling, 2018). He blames this on three factors: fake nostalgia (a misremembering of the past as being better than it actually was); selective reporting by journalists (e.g. emphasising negative stories in accordance with traditional news values that prioritise exposure of suffering, corruption and wrongdoing in accordance with traditional news values); and a feeling that it is somehow inappropriate to talk about minor improvements during crises. The spread of disinformation often preys on and manipulates these beliefs, particularly where crises, conflicts and natural disasters are concerned. While Rosling encourages the notion of public education as a countermeasure, it remains a research gap to understand how effective this strategy is, especially given Rosling's own findings.

The "Ticks or It Didn't happen" report by Witness (Witness Media Lab, 2019) focuses on responses to disinformation from a primarily ethical viewpoint. Taking one of the core technologies for tracking image integrity ('point-of-capture' approaches at a camera level), the report reviews 14 dilemmas that are relevant since authenticity infrastructure is considered as a response to misinformation, disinformation and media manipulation. These dilemmas include technical dilemmas around access, as well as privacy, surveillance, government co-option, and concerns about setting overly-simplistic or hard-to-assess markers of credibility. The lens of the report is to use the framing of Collingridge's dilemmas (Collingridge, 1980) on the capacity to influence technological systems - and the challenge of doing that early enough to ensure they reflect human rights values, or risking being excluded once they are at scale. This lens is, however, also applicable to a range of technological approaches to disinformation, that may or may not prioritise freedom of expression or other human rights issues.

[25] Cognitive miserliness is the notion that we prefer to make easy decisions that align with our preconceptions, and may forget details (such as that the information had previously been debunked) https://firstdraftnews.org/latest/the-psychology-of-misinformation-why-were-vulnerable/

3.2.2 Vector focus

Alongside notions of persuasion and countering false beliefs, responses to disinformation also need to take into account the nature of the disinformation, and at whom it is aimed (as discussed above), but also the role of the conveyancing mechanism, or vector. These serve as intermediaries between the production and consumption of disinformation, enabling its circulation in various ways and at various scales. Knowledge about the patterns in this part of the cycle is critical for informing responses not only within transmission, but also in regard to strategies that target the initial production and subsequent consumption of disinformation.

There are three main mechanisms by which false content may be conveyed. First, disinformation may aim to disrupt or leverage the news media as a way to indirectly reach its targets, whether these be state or non-state actors. Captured media, compromised journalists, or weak capacities for verification constitute vulnerabilities that are exploited. Alternatively, disinformation may appear as a strategised (and often, but not necessarily, automated) exploitation and/or gaming of an internet platform to reach the public (i.e. targeting in part the nature of the business model and its reach). In other cases, disinformation is aimed primarily at the public for the purpose of onward dissemination, relying on its potential to trigger virality, using third parties to serve as peer-to-peer intermediaries to reach a bigger audience. In each case, responses need to target primarily the relevant mechanism (media, internet company, and public respectively).

3.2.3 Defending public values in a 'platform society'

While the news media and the public may serve as vectors for disinformation, this chapter now considers in more detail research into the role of internet communications companies (often referred to as 'platforms') as conduits, amplifiers and atomisers for disinformation. The rise of digital technologies has led to the increasing importance of data, with these companies emerging as new bastions of control and profit, having the facility to capture and manipulate enormous volumes of content- and, potentially, audiences. This in turn has led to the rise of dominant players (Srnicek, 2017), and it has important ramifications for the production, dissemination, and consumption of information and its reliability. An initiative by the NGO Public Knowledge, operating as https://misinfotrackingreport.com/, keeps pace with the policies and practices of a number of companies dealing with the challenges. Civil society movement Avaaz tracks the visible manifestations of disinformation narratives on specific themes, evaluating the performance of the companies in combatting such content.[26]

To some extent, the business models of digital platforms make them vulnerable as the conduits of disinformation, but there is also an argument that they are actually de facto enablers, or accomplices who turn a blind eye to the issue (Gillespie, 2017). Gillespie suggests a definition for the modern concept of (internet) platform as: "an architecture from which to speak or act, like a train platform or a political stage." However, like a growing number of researchers, he shuns the notion of 'platforms' because it tends to underplay the particular role of the companies involved. Gillespie points out that in reality, online platforms are not flat, open, passive spaces, but "intricate and multi-layered landscapes, with complex features above and dense warrens below." This suggests that such a complex structure influences how content is transmitted, and in ways that are not immediately open or straightforward. Instead, the nature of the online content that

[26] https://secure.avaaz.org/campaign/en/disinfo_hub/

users receive is shaped by algorithms, and can also change dramatically at the behest of those who have control of the design of the platform. The business model can enable bots and trolls to lurk beneath the surface and strike at unsuspecting victims or types of information, as well as enforce systemic biases such as decisions on what is allowed or not, and what might be a trending topic. That is one reason why the word 'platforms' is used sparingly in this report - instead, wherever feasible, the term 'internet communications companies' is used in preference.

Relevant to this issue are the financial gains to be made through the analysis of enormous amounts of data made available to companies which enable transmission or discovery of content. Zuboff (2019) has assessed how engagement is required from users, in order to produce this data, which is then monetised in the form of opportunities that are sold due to their ability to shape what she calls "behavioural futures". Reports from Ranking Digital Rights highlight that this business model leads to particular kinds of content becoming more widespread, including disinformation. By prioritising such content and recommending similar content, disinformation becomes increasingly linked with revenue for both platforms and the content providers, and the problem becomes circular (Maréchal & Biddle, 2020; Maréchal et al., 2020).

The book "The platform society: Public values in a connective world" (van Dijck et al., 2018) also offers an in-depth analysis of the role of these companies in shaping modern society. It focuses on public values in a world where social interaction is increasingly carried out on digital platforms, and investigates how these values might be safeguarded. Until recently, most companies have tended to evade acceptance of the social obligations related to their position as intermediaries of content, although this is beginning to change as pressure is put on them by authorities, especially European policymakers. While some companies have encouraged research into disinformation, there is reluctance to make their data available for this purpose. For example, Facebook has announced $2m for research into "Misinformation and Polarisation" with the proviso that "No data (Facebook, Messenger, Instagram, WhatsApp, etc.) will be provided to award recipients".

Another area ripe for further research in reference to the role of the internet communications companies' in combatting disinformation is the exploitation of 'data voids' (Golebiewski & Boyd 2019). Research being conducted at the time of writing, as part of a partnership between First Draft and researchers from the University of Sheffield, identified the particular problem posed by data voids during the COVID-19 pandemic. They found that when people searched for answers to questions about the causes, symptoms and treatments for coronavirus, the void created by the absence of verifiable answers to these questions (in part a product of the genuine scientific uncertainty associated with the onset a new virus; sometimes because of manipulated disclosure by authorities of statistical data) lent itself to exploitation by disinformation agents who filled the gap with spurious content: "If more speculation or misinformation exists around these terms than credible facts, then search engines often present that to people who, in the midst of a pandemic, may be in a desperate moment. This can lead to confusion, conspiracy theories, self-medication, stockpiling and overdoses." (Shane 2020) On the basis of preliminary findings, and recognising the role that social media sites now play as de facto search engines, the researchers called for a 'Google Trends' like tool to be developed for application to a range of social media sites including Facebook, Twitter, Instagram and Reddit, to enable easier and more transparent identification of disinformation being surfaced by such search activity.

The intersection between internet companies and news media companies as vectors for false content has also attracted some analysis. In particular, this highlights tensions between journalism and internet communication companies with respect to curatorial

efforts to counter disinformation and its viral distribution, the purveyors of which frequently target journalists and news publishers. These tensions have their roots in the 'frenemy' status of the relationship between these companies and news publishers (Ressa, 2019), which has been exacerbated by the collapse of traditional news business models, the erosion of historic gate-keeping roles, and the rise of 'platform power' (Bell & Owen, 2017).

The escalation of digital disinformation in the context of journalism's dependency on these social media networks for content distribution and engagement, and the platforms' encouragement of such dependency, have led to the phenomenon of 'platform capture'. Other examples of 'platform capture' include the ways in which efforts to curtail disinformation can backfire, such as WhatsApp's change in terms of service in 2019 which negatively affected the media's ability to use the technology to counter disinformation (Posetti et al., 2019b).

Traditional journalism commits to a set of news values (Galtung and Ruge, 1965) that include accuracy, verification, and public interest, but this is potentially orthogonal to the values of digital platforms which typically include innovation and peer-to-peer connectivity (Wilding et al., 2018), not to mention monetisation at the expense of editorial standards. As Foer (2017) indicates, dependence of the news media on the values of the digital platforms, means that their intensified quest to go viral risks superseding the quest for truth. This problem is further exacerbated by algorithms for the optimisation, dissemination and even production of news (Wilding et al., 2018) as well as search engine optimisation.[27] In addition, audience engagement has become a core driver, resulting in a change in news production towards a "softer" form of news (Hanusch, 2017) that is shorter, more visual, and more emotive (Kalogeropoulos et al., 2016). Added to this, 'content farms' are producing or recycling questionable low-quality content with dubious factuality but which are optimised for engagement.

The digital transformation of journalism is ongoing - change is now regarded as a perpetual - therefore, it is important that research keeps pace with the associated challenges and opportunities relevant to the production, dissemination and amplification of disinformation in the 21st century news ecosystem (Ireton & Posetti 2018).

An assessment of the internet and news media vectors, and the relationship between them, are discussed in detail in Chapter 6.1 of this report.

3.2.4 Policy-driven approaches to studying disinformation

The COVID-19 crisis prompted a range of studies with a view to developing policy responses, including by UNESCO (Posetti & Bontcheva, 2020a and 2020b) and the OECD (2020). The OECD study used the Wardle and Derakhshan (2017) framework to identify four governance responses to disinformation: identifying and debunking; civic and media initiatives; communications strategies; and regulatory measures. Particular attention was focused on public communication with the message that "Strategic and transparent communication should be among the first lines of action for public institutions at all levels".

The LSE's *Tackling the Information Crisis report* (LSE, 2018) explains how changes in the UK media system have resulted in what it calls an information crisis. It depicts this as being

[27] https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/08/Follow-the-Money-3-Aug.pdf

manifested in 'five giant evils' among the UK public – confusion, cynicism, fragmentation, irresponsibility and apathy. It also summarises a number of UK policy responses, including UK parliamentary inquiries; UK government initiatives including among other things the Digital Charter (UK DCMS & Rt Hon Matt Hancock, 2018b), a white paper on new laws to make social media safer (UK DCMS, Home Office, Rt Hon Matt Hancock & Rt Hon Sajid Javid, 2018a), and the new DSTL Artificial Intelligence Lab in Porton Down, whose remit includes "countering fake news" (UK MOD et al., 2018); institutional responses such as those by Ofcom (2018b) and the Commission on Fake News, and the teaching of critical literacy skills in schools (National Literacy Trust, 2018). While the report provides a detailed coverage of policy responses to disinformation, it focuses primarily on recommendations and recent initiatives, but research is still needed on analysing the outcome and impact of these.

Launched in November 2018, the Information Warfare Working Group[28] at Stanford University, comprised of an interdisciplinary group of researchers at the Center for International Security and Cooperation at the Freeman Spogli Institute and the Hoover Institution, aims to "advance our understanding of the psychological, organizational, legal, technical, and information security aspects of information warfare", working towards producing a set of policy recommendations for countering foreign disinformation threats. They have so far produced a number of white papers and reports. The work comprises research from many different disciplines and foci, while at the same time it focuses rather narrowly on political aspects of disinformation in the U.S..

Other important resources at the European policy level include a study commissioned by the European Parliamentary Research Service investigating the effects of disinformation initiatives on freedom of expression and media pluralism (Marsden & Meyer, 2019), as well as the work of the High Level Expert Group (HLEG) on 'Fake News' and Online Disinformation (Buning et al., 2018).

The first of these reports examines the tradeoffs between the application of automated (AI) techniques to counter disinformation, focusing mainly on ways in which EU legislation can be used to drive the design of these technologies in a way that does not restrict freedom of expression unnecessarily, and which maximises transparency and accountability. It thus focuses primarily on technological and legislative responses to disinformation, and raises concerns over the nature of current legislation that might restrict freedom of expression, concluding that there is a lack of policy research in this area, and that single solutions, particularly those which focus primarily on technological responses, are insufficient. In a similar vein, the HLEG report provides a policy-driven perspective on disinformation, advising against simplistic solutions and encouraging holistic solutions promoting maximum transparency, information literacy and empowerment, and suggesting a mixture of short- and long-term actions.

Both these reports thus focus specifically on European policy issues, and thus do not consider how this might be translated beyond these boundaries. Indeed, a major research gap in all the existing policy-driven reports is that each proposes their own set of strategies but it is unclear how to proceed from this to an overarching set of responses, even though disinformation clearly does not respect geo-political boundaries.

A group of experts from the University of Pennsylvania have produced a report titled "*Freedom and accountability. A transatlantic framework for moderating speech online*" (Annenberg Public Policy Center, 2020). This document states that: "Through a freedom-

[28] https://cisac.fsi.stanford.edu/content/information-warfare-working-group

of-expression lens, we analyzed a host of company practices and specific laws and policy proposals, gathering best practices from these deep dives to provide thoughtful contributions to regulatory framework discussions underway in Europe and North America." To deal with online problems, including disinformation, the report proposes that States should regulate internet companies on the basis of compulsory transparency provisions, and that there is also regulatory oversight to "hold platforms to their promises". For the internet companies themselves, the report suggests a three-tier disclosure structure, effective redress mechanisms, and prioritisation of addressing online behaviour by "bad actors" before addressing content itself.

3.2.5 Practice-relevant studies and resources

The Digital News Report[29] of 40 markets from the Reuters Institute for the Study of Journalism documents the role that internet companies are now playing in the distribution of both information and what the report's authors call "misinformation". It points out that audiences can "also arrive at misinformation (as they arrive at much else) side-ways via search engines, social media, or other forms of distributed discovery". The 2018 report in the series examined variations in exposure and concern, and different beliefs about remedies to false content online. The Institute has also researched types, sources, and claims of COVID-19 misinformation (Brennan et al., 2020), and mapped disinformation responses from three Global South news organisations re-conceptualising themselves as 'frontline defenders' in the 'disinformation war' (Posetti et al., 2019a; Posetti et al., 2019b). These latter reports identify enhanced methods of investigative reporting (including big data and network analysis), advanced audience engagement techniques (such as collaborative responses to surfacing and debunking disinformation), and 'advocacy' or 'activist' models of journalism (that involve actively campaigning against disinformation vectors, or providing digital media literacy training to their communities) as methods of responding to the disinformation crisis.

Jigsaw (an arm of Google) has produced what they term a **visualisation of disinformation campaigns** around the world, supporting their theory that "understanding how disinformation campaigns operate is one of the first steps to countering them online".[30] They state that this visualisation is based on the Atlantic Council's DFRLab research and reflects their perspectives in the characterisation. Additionally, they note that their visualisation is primarily based on open source, English-language press reporting of campaigns which appear to target the West. These kinds of visualisation provide an interesting overview, despite geographic limitation, but risk conflating very different kinds of disinformation.

A noteworthy set of practical resources pertaining to disinformation includes some of those discussed in chapters 7.1 (focused on normative and ethical responses) and 7.3 (educational responses), which not only support practical skills, but also investigate underlying theories and trends. The UNESCO handbook *Journalism, 'Fake News' and Disinformation* (Ireton & Posetti, 2018), is a research-based educational resource aimed at journalists and news organisations confronting disinformation, with an emphasis on freedom of expression issues. In addition to its role as a set of resources to support journalism education, it also explores the nature of journalism with respect to trust, as well as the structural challenges that have enabled viral disinformation to flourish, and the conduits of information disorder such as digital technology and social media, and it

29 http://www.digitalnewsreport.org/
30 https://jigsaw.google.com/the-current/disinformation/dataviz/

describes the targeting of journalists and their sources in the context of disinformation campaigns. The book offers a framework for understanding independent, critical journalism as a mechanism for combatting disinformation. It also provides models for responding innovatively to the challenges of disinformation as they impact on journalism and audiences. Among other resources of this kind is the *Verification Handbook for Disinformation and Media Manipulation* produced by the European Journalism Centre (Silverman, 2020).

Examples of resources focusing on the public include the UNESCO MIL Digital Toolkit[31] comprising MOOCs on Media and Information Literacy in several languages, and the International Center for Journalists' (ICFJ) learning module on the history of disinformation (Posetti & Matthews, 2018). One important gap in a number of these toolkits and programmes is a focus on the wider representation of 'data', including privacy and profiling issues, and more generally how data is collected and used by online platform providers, as discussed earlier in this chapter. The experience of the NGO 5Rights, for example, has shown that when children understand these concepts, their overall information literacy also improves. However, many skills-based approaches to countering disinformation only focus on the basic concepts of verification of the immediate sources without considering these wider foundational aspects.[32]

Finally, a handbook for government communicators on countering information influence activities has been produced by the Swedish Civil Contingencies Agency (MSB, 2020).

31 https://en.unesco.org/MILCLICKS
32 https://5rightsfoundation.com/uploads/digital-childhood---final-report.pdf

3.3 Current research gaps

As has been indicated, there is a plethora of research on disinformation and approaches to countering it, both from a theoretical and practical standpoint. However, there is an apparent disconnect between academic research, journalistic investigations, and studies commissioned by civil society and intergovernmental organisations. Additionally, actual collaboration between these sectors appears to be infrequent. The initiatives and publications mentioned have been produced in an ad-hoc manner, and are disparately located, making it difficult to track, analyse, and synthesise them in a coherent way. For example, cross-institutional study of the relationship between the technological/business logic and the realm of company and state policies is still weak, as will be discussed further in chapter 6.2.

The impact of most of the responses counteracting disinformation has also not been studied sufficiently. While some research has investigated which groups (such as elderly people) are particularly susceptible to both believing and sharing disinformation (Carey et al., 2020; Guess et al., 2019), there have been few responses directly aimed at vulnerable groups, and there is a dearth of empirical assessment of these, with exceptions like Humprecht et al. (2020), although with limited geographical focus. Linked to this, methods of countering disinformation have also not sufficiently covered notions of group allegiance and distrust in authority, which require a different outlook and more fundamental issues to be addressed.

Finally, while there is a growing body of research, software, training and resource development focused on tackling disinformation, there is a comparative absence of that which focuses on disinformation in the light of human rights, freedom of expression, and the burgeoning access to - and use of - broadband technology worldwide.

Below is a further analysis covering some particular areas where important gaps have been identified.

Addressing distinctions and connections between realms of disinformation

In terms of frameworks, much published research does not make a clear distinction between novel kinds of disinformation (for example, deepfakes) and those with much older histories (such as notions of information influence, which overlap with disinformation as discussed above). Others apply only in specific contexts, such as political disinformation, or may have limited applicability to non-Western nations (Brooking et al., 2020). A number of frameworks also view disinformation not only in a political light, but also focus primarily upon foreign influence, and thus do not address the numerous issues related to domestic disinformation, such as that pertaining to health crises, issues of migration, and disaster communications.

On the other hand, there are separate specific studies around such issues, as witnessed by the latest efforts to map disinformation around the COVID-19 pandemic and to implement counter-strategies, discussed in more detail in the following chapters. In general, the effect of the pandemic has been to ramp up public awareness of disinformation, and educational efforts promoted by both state and non-state actors (governments, internet communications companies, media companies, etc.). COVID-19

has provided a clear case where the effects and harms of disinformation can be easily seen, thereby elevating its importance and dangers in the public's eyes, and may lead to increased research, such as the initiative of the World Health Organisation (WHO) to explore an interdisciplinary field of "infodemiology" study, which has relevance to fields outside of health.[33]

Data availability for research

In terms of understanding the nature of disinformation, its dissemination and counter-activities, the issue of the lack of transparency of algorithms behind social media platforms and issues with access to their data is a serious hindrance, as discussed in chapter 4.2. Quantification of disinformation online relies on selective disclosure by the companies and what is contained in their transparency reports, without researchers having access to original data.

There is evidence, from external studies, about instances of disinformation pieced together through content analysis techniques. One snapshot study said it found that one in four popular YouTube coronavirus videos contained misinformation.[34] This research analysed 69 of the most widely-viewed English language videos from a single day in March 2020 and found 19 contained non-factual information, garnering more than 62 million views. In another study, an analysis of more than 1300 Facebook pages with nearly 100 million followers produced a network map showing that while anti-vaccine pages have fewer followers than pro-vaccine pages, they are more clustered and faster growing, and increasingly more connected to other pages.[35]

Such findings signal the importance of assessing patterns of disinformation online, and they also show what can be done even without data disclosed by the internet companies.

Nevertheless, most research into disinformation is limited by being conducted without access to the complete data sets from the internet communications companies. This leads to a lack of depth in their analysis, and studies are also typically carried out only on a selected platform (frequently Twitter with its volume of open and public data), rather than cross-platform. Messaging apps are rarely considered due to their closed nature. Social media companies present a number of obstacles to independent research by cutting access to APIs by which researchers can collect relevant data, mirroring to some extent the problems with search engine research, where only those with direct relationships with the major search companies can work effectively (Walker et al., 2019). For instance, it is hard to know specifics when users or messages are removed by the provider (or when the user retracts the information themselves). While these platforms do offer a selected group of academic researchers to access such data via research grants[36] by means of tools such as Crowdtangle[37], at the time of writing this was limited in scope and included restrictions on the kinds of research that could be done. In the light of COVID-19, Crowdtangle launched (in March 2020) more than 100 publicly available LiveDisplays enabling researchers to investigate issues such as the spread of information about the pandemic on social media, nevertheless this still provides only a restricted set of data.

33 https://www.who.int/docs/default-source/epi-win/infodemic-management/infodemiology-scientific-conference-booklet.pdf?sfvrsn=179de76a_4
34 https://www.nbcnews.com/health/health-news/live-blog/2020-05-13-coronavirus-news-n1205916/ncrd1206486#liveBlogHeader
35 https://www.nature.com/articles/s41586-020-2281-1
36 https://about.fb.com/news/2019/04/election-research-grants/
37 https://www.crowdtangle.com/

The report of the Annenberg Public Policy Centre (2020), cited above, argues that transparency enables governments to develop evidence-based policies for oversight of internet companies, and pushes firms to examine problems they would not otherwise address, and thus empowers citizens. This insight points to the value of companies providing much greater access to data. Companies are understandably sensitive about providing data for reasons of commercial secrecy as well as avoiding data compromises, as occurred during the Cambridge Analytica experience. Against this background, MacCarthy (2020) has proposed the nuance of a tiered model for access to company data, distinguishing different levels that could be availed to the public, vetted researchers, and regulators.

The consumption and response to disinformation

Studies in **user behaviour and perception** are still lacking, not least in regard to the relationship between the impacts of disinformation and of news. For example, even when faced with a diverse selection, people tend to choose news articles that are most aligned with their own beliefs (Kelly Garrett, 2009) - through user-driven customisation or selective exposure, reinforced by predictive algorithms. Nevertheless, little work has been carried out on assessing its actual effect. This has important ramifications for disinformation with respect to issues of propaganda or dangerous health-related beliefs such as those promoted by anti-vaccination supporters. The implications of such selective exposure are of increasing concern, since they can enhance social fragmentation, mirroring or amplifying enduring cleavages, thereby also reinforcing pre-existing opinions and perceptual biases. The correlation between exposure to misinformation and effects on offline behaviour also requires further investigation, such as the relationship between misinformation, fear, panic, and unselfish and irrational behaviour (see e.g. Osmundsen et al., 2020).

Competing notions currently exist around the extent and effect of exposure to different viewpoints on one's ideological perspectives. On the one hand, the increasing use of social media and personalised news acts as a 'filter bubble' or 'echo chamber', reinforcing existing beliefs and increasing ideological segregation. However, there is a growing body of empirical research arguing that the effect of filter bubbles has been overstated (e.g. Dubois & Blank, 2018; Guess et al., 2018a), and that only a small subset of people tend to have heavily skewed media consumption (Gentzkow & Shapiro, 2011; Flaxman et al., 2016), something which extends also to misinformation (Guess et al., 2018b). Others posit that the increasing availability of information, coupled with the consequent greater diversity of the information consumed, actually widens the range of news sources to which people are exposed (Fletcher & Nielsen, 2018). Another study showed that even users of very different political backgrounds were typically exposed to very similar sets of political news (Nechushtai & Lewis, 2018), contradicting theories about the filter bubble effects of news personalisation.

What is unclear is what effect widening the exposure to different viewpoints might have on issues of ideological partisanship. Understanding and measuring ideological diversity from big social data, and the influences on ideological perspectives that might be brought about by exposure to such diversity, would all lead to improved understanding of the effect of disinformation and counter-content such as fact-checking and verified journalistic news. Large-scale user studies would be needed in order to better understand how people evaluate the truth and reliability of information – both from a practical perspective and from a psychological perspective. Similarly, studies targeting users' behaviour in relation to engagement with, and redistribution of, credible, verified

information - such as that produced by independent news publishers and journalists - could provide insight.

Several attempts have been made to mitigate the effect of bias in information systems to support an unfiltered process of opinion formation. Some have focused on making users aware of bias by providing alerts (Epstein & Robertson, 2015), visualising reading behaviour and bias (Munson et al., 2013), or pointing to web pages with different opinions from the current one. Others rely on visualisations to support diversity in web activities (Graells-Garrido et al., 2016), recommendation systems (Tsai and Brusilovsky, 2018), and search results (Verberne, 2018). Some focus on algorithm transparency by explaining how filtering works and enabling the user to control the algorithm and thus their filter bubbles (Nagulendra & Vassileva, 2014). Others try to break potential filter bubbles through software design and user interfaces (Bozdag, E., & van den Hoven, J., 2015). However, success in all of these approaches is rather limited (Faridani, 2010; Liao & Fu, 2013), and more studies are clearly needed to better understand online news consumption patterns and habits, such as how people navigate the constantly changing environments to select which news they decide to read (Swart et al., 2017).

The changing technological and institutional infoscape

It can be noted that many of the responses to disinformation described in this report are still quite new and not yet widely implemented. This may be because the technologies are still being developed or improved, because they are only adopted by a small minority, or for other reasons such as legal and ethical issues which need to be resolved. For example, when credibility and labelling approaches are not widely used, this not only has clear limitations on their effectiveness, but also on the understanding of their potential. It is simply not known if they will be successful until they are rolled out more widely. There are also potentially serious implications if they are applied at scale, as detailed in the 'Ticks or It Didn't Happen' report by Witness (Witness Media Lab, 2019). This illustrates Collingridge's dilemma (Collingridge, 1980), which essentially posits that the social consequences of technology often cannot be predicted until the technology has already been developed, at which point it is often too late, or at least much more difficult to change. Neither Collingridge nor the Witness report suggest that these challenges cannot be overcome, but focus on early consideration of scenarios, as well as flexibility of approach in order to deal with them.

Related to this, evaluation of many of the technologies proposed to counter disinformation is still lacking, and furthermore little discussed. It is not always even clear how effective some of the methodologies are in principle, such as the notion of fact checking, since research has shown that the reach of fact-checked material is often very different from the reach of the disinformation itself, and indeed, instances of a "backfire effect" have been witnessed where corrections can sometimes even increase misperceptions (Nyhan, 2012). More research could help in evaluating the effect not only of the technologies but also their underlying theories of change, which may be based on false or misguided assumptions. Further discussion of this is presented in Section 4.1.

International representativeness in research

The Global South in particular has typically been under-represented in terms of research focus. Examples include Chaturvedi's study of India (Chaturvedi, 2016); Kaur et al.'s study of Asia and the Pacific (Kaur et al., 2018); recent reports of joint research between FullFact, Chequeado and Africacheck focusing on Argentina, South Africa and Nigeria[38], and the Oxtech report on anti-disinformation initiatives, which uses examples from 19 countries on four continents.[39] Reports from a policymaker's perspective, in particular, are almost exclusively focused on Europe and North America. That is a clear gap that this study aims to address, partly in the hope that it will trigger investment in future action-oriented research.

All this highlights the value of a large-scale global study such as this one, which collates the multiplicity of disinformation responses from a variety of perspectives, and incorporates the needs and challenges of culturally distinct geographical regions.

Human-rights dimension

Few conceptual frameworks or other literature really focus on the critical problem of ensuring a balance between protecting freedom of expression and upholding notions of truth, against disinformation, although this is connected implicitly with some of the discussions in this chapter around the internet communications companies, as well as around journalistic integrity. Meanwhile, regulating speech on social media in an attempt to prevent disinformation clearly has ethical and policy implications that intersect with freedom of expression, as does the passage of legislation creating 'fake news' laws that represents a significant threat to press freedom. The EU Code of Practice on Disinformation (European Commission, 2018c) has recently been criticised for theoretically allowing, and even incentivising, restrictions on the freedom of speech that are claimed to be technically lawful (Kuczerawy, 2019). Kuczerawy voices concerns that enlisting private platforms to suppress certain online content that is not illegal may have unintended consequences, and argues that it is difficult to "factually assess the impact of the Code on the exercise of the right to freedom of expression". In countries outside the EU, where less stringent regulations may apply, there is the potential for greater concerns of this nature. These issues are discussed more fully later in this report, in particular in the discussions of legislative responses to disinformation in Chapter 5.1, as well as in the discussions of policy responses in Chapter 5.2, since both these kinds of responses must deal with this exact issue.

[38] https://fullfact.org/research/

[39] https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/A-Report-of-Anti-Disinformation-Initiatives

3.4 Novel contributions of this study

Having situated this study within the context of existing theoretical frameworks and previous research, and having identified the gaps in current research on the topic of disinformation, this section highlights the specific novel contributions presented here.

Firstly, this study has sought to adopt a global focus, while many of the reports cited above have largely focused on particular countries or continents and a great amount of research has centred on the UK, U.S. and/or European situations. This partly reflects the fact that these geographical regions are highly active in responses to disinformation, and that they represent the location of the majority of researchers and funding for investigating the topic. Further, the dominant disinformation sources under examination in other reports have been limited to English language content.

By contrast, this report has sought to focus on issues and initiatives worldwide, including those from Africa, Australia, Central and Eastern Europe, Latin America and Asia. For example, this has helped reveal that some journalistic responses to disinformation rely on having certain technological requirements, or are difficult to adopt for those in conflict situations (such as when reporters need to maintain anonymity and cannot use certain point-of-capture tools for photos and videos as a result). Below, we discuss how and why particular responses may be difficult for actors in certain countries and situations, which are not necessarily considered by those in Western Europe and the U.S..

The authors of this report are of diverse ethnic and regional backgrounds, they speak a variety of languages and they possess specific knowledge about situations in different parts of the world. They also come from a range of disciplinary backgrounds. The research team includes members from both academia and industry, with a mixture of computer scientists, journalists, social scientists (including those with a journalism studies and political science background), and specialists in international human rights with an emphasis on freedom of expression. This leads to an approach which addresses a range of perspectives and is closely tied to both practice and impact. There is thus also a focus on technical responses such as the use of AI, in addition to educational responses, responses from the journalism sector, and responses from the industrial technology sector.

This report is also novel because it puts the main focus specifically on **responses** to disinformation. As discussed above, other notable reports focus on dilemmas (Witness Media Lab, 2019), policy implications (e.g. LSE, 2018; Annenberg Public Policy Center, 2020), political implications (e.g., Marsden & Meyer, 2019; Pamment et al., 2018), and significance for (as well as responses from) journalism (Ireton & Posetti, 2018; Posetti et al., 2019a). Furthermore, this report addresses the entire spectrum of disinformation responses, rather than focusing on a specific type such as political disinformation campaigns (Brooking et al., 2020) or issues with access to company data and how this affects academic research (Walker et al., 2019).

A further novel angle of this study is that the problem of disinformation is systematically addressed in the light of freedom of expression challenges, with implications for press freedom such as in legislative responses, among others.

The typology of responses that this study has developed also breaks down the problem of disinformation in a new way. It examines each response from a variety of perspectives, looking beyond the what and how to issues such as "Who is funding these responses (and the implications thereof)?", "What are the strengths and weaknesses of them?", and "What is the theory of change on which they are based?" This approach provides additional insight into the assumptions upon which the responses rest, and the extent to which they integrate monitoring and evaluation into their activities.