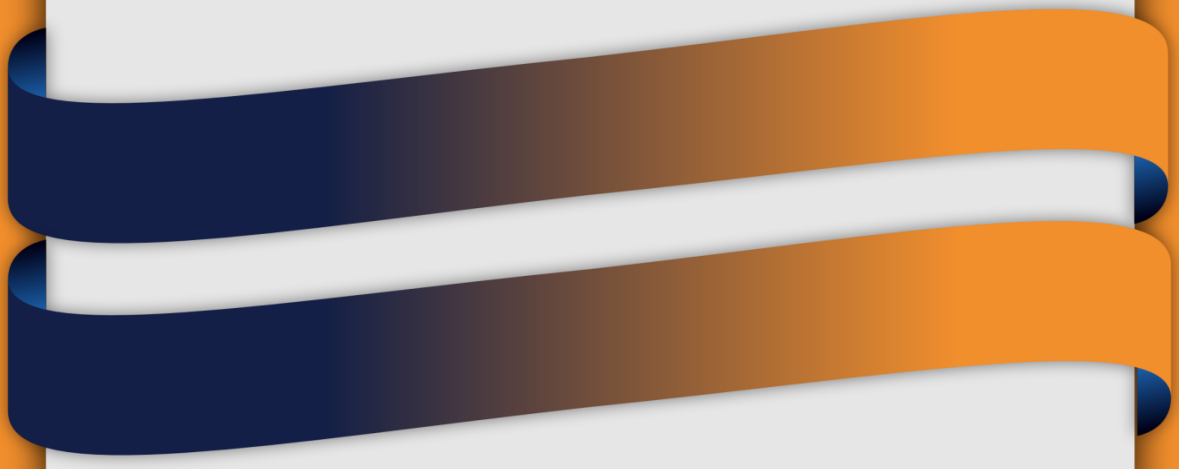
PANEL ON MISINFORMATION AND FREE SPEECH IN MODERN DEMOCRATIC SOCIETIES

FIDE OXFORD/22

JANUARY **2023**



Fide
Law | Economy | Technology

Relevant Information

This document summarises the key point of one of the round tables held at the Fide Foundation 2nd International Congress at Oxford, on Nationalism, Populism and Identities: Contemporary Challenges. The key topic was the Impact of nationalism and populism at the national level.

The panel was comprised of **Dr Talita Dias,** Shaw Foundation Junior Research Fellow in Law, Jesus College, Oxford, **Teresa Rodríguez de las Heras Ballell,** Professor of Commercial Law at the Carlos III University of Madrid. Sir Roy Goode Scholar at UNIDROIT. Academic Advisor to Fide. *(Leader of the Group) y* **Guillermo Serrano,** Public Policy Manager, Spain & Portugal, Meta. Currently Head of Government Relations in Trainline España

About the Fide Foundation

The Fide Foundation is a legal-economic think-tank based in Spain, committed to involving the civil society in all major legal and economic developments in Spain, the EU and abroad.

Website: thinkfide.com

TABLE OF CONTENTS:

N. PANEL ON MISINFORMATION AND FREE SPEECH IN MODERN DEMOCRATIC SOCIETIES

ABSTRACT

Dis- and misinformation are challenging phenomena in today's digital society. The viral dissemination of false information online can undermine public debate and erode trust both in science and democratic institutions. In the following document, we approach these issues from three angles: 1) the balance it entails between conflicting fundamental rights; 2) the complexity of the problem, requiring multistakeholder responses with a focus on how digital platforms address this phenomenon; and 3) its impact on the authoritative role of science in democratic societies. We conclude that, while an intermediary liability regime for online platforms is inappropriate, several tools exist and are at the disposal of companies, states and other actors in tackling misinformation and disinformation campaigns.

INTRODUCTION

The pervasive phenomenon of dis/misinformation in contemporary societies presents a difficult challenge, as it erodes the quality and health of public debate and democratic institutions, and brings about an unstable, delicate balance of conflicting rights and interests. This panel aims to provide a critical analysis of the phenomenon of dis/misinformation (including 'fake news' and other information operations), its spread (conceptualisation, underlying motives, dynamics, taxonomy, etc.) and its interactions with the pillars and values of democratic societies. We also assess the scale and the relevance of the problem, anticipate solutions, and contextualise the conclusions.

The panel has approached the phenomenon from three main angles:

First- the balance between conflicting fundamental rights

Misinformation and disinformation pose a challenge to modern democratic societies, as they erode the quality of the public debate and trust in its institutions. Free speech, on the other hand, remains one of the very pillars underpinning democracy itself. How can the negative effects of dis/misinformation be prevented whilst freedom of speech is preserved? How to strike the right balance between conflicting fundamental rights?

Second- dis-/misinformation as a complex phenomenon to be decoded

Dis/misinformation is a highly complex and multi-layered phenomenon that requires the cooperation and involvement of many actors, including governments and institutions, academics, journalists, digital platforms, civil society and/or individual citizens themselves. How do we allocate not just the legitimacy to act, but also responsibilities amongst them? What role should each of these actors play?

Third- the impact of dis/misinformation on the authoritative role of science in democratic societies

Looking at specific cases, organised misinformation and disinformation campaigns surrounding the COVID-19 pandemic contributed to creating misconceptions and to the spread of inaccurate information. This had, and continues to have, an impact on individual and overall public health, as well as eroding citizens' collaboration and trust in national and international public health institutions strategies to contain the pandemic. How can the negative impact of health misinformation be prevented? To what extent has dis/misinformation undermined the authoritative nature of science?

KEY POINTS

The complexity of the phenomenon and its permanently changing character invites us to devise an analytical and theoretical framework to capture the key issues. The proposed framework is based on four main pillars, as described below:

I. Concepts, Definitions, and Delimitation of Scope

The mis/disinformation binomial conceals a polyhedric, complex, and evolving phenomenon. An initial conceptualising effort is essential decisive. The aims are:

- To differentiate **misinformation** and **disinformation** (by means of intent, deliberateness, inauthentic behaviour, the promoting actors, propagators, context, etc.). As of yet, there is no formal definition of such types of operations under international law. A useful typology, though, is one based on the authors' intentions and the verifiability of the information deployed:

(1) *Misinformation* –false information that is often shared unintentionally.

(2) *Disinformation* – coordinated efforts that aim to manipulate or corrupt the public debate to fulfil a strategic goal. It is characterized by two features: inauthenticity and coordination. Disinformation campaigns are often named "Influence Operations" or "Information Operations".

- To **distinguish** these from content that consists of **opinions**, and therefore cannot be deemed true or false, and also from content that can be deemed as illegal or harmful. Falsity may arise from a diversity of situations (context, chronology, facts, etc.).

- To provide a **taxonomy** of 'fake news' as a mere buzzword that embraces a variety of mis/disinformation strategies or situations, on the basis of selected factors (facts, falsity, misleading potential, context, media, etc.) and to properly identify the distinctive characteristic of **information operations (disinformation)** and misinformation.

- To identify the new features of the phenomenon in **the digital society** as compared to the pre-digital use of propaganda, weaponized disinformation campaigns, and biased information: that is, the **viral**

nature of 'fake news', the role of **social media and digital platforms**, and the **absence of authoritative references.**

On 'fake news' and its potential harm: the role of civil liability rules - Teresa Rodríguez de las Heras Ballell

The term 'fake news' has become extraordinarily popular not just in describing various forms of misinformation and disinformation, but also to denote purely illegal content, defamation, parody, or simply offensive content. As a consequence, 'fake news' is a useful term to direct attention towards a well-identified social problem, although the concept is, in essence, vague, imprecise, and to a certain extent, too confusing to be employed in legal analysis. On the one hand, the "fake news" phenomenon certainly comprises more than news. It encompasses any visual, graphic, or textual content produced and disseminated on a digital format that is likely to misinform. On the other hand, the 'fake news' label is also used to tag a wide array of mis- and disinformation types, including manipulated content, false content, misleading content or fabricated content. With such imprecision, even if it describes an apprehensible reality, the term is unsuitable for delimiting the scope of application of any regulatory action.

Should the delimitation of the scope be approached from the perspective of civil liability, a categorisation based on types of potential harm deriving from the content at stake becomes relevant. Harm caused by digital content can be varied in nature (moral, reputational, patrimonial, or even indirectly physical or personal) and may differ in extent. Whereas some digital content is likely to cause damage to identified persons (either natural persons or moral ones), other content does simply generate diffuse or collective harm. In the latter case, despite the severity of the harm and the ampleness of the negative impact, no specific victims can be singled out. Even proper 'fake news' in a strict definition does very frequently fall under this last category.

The spread of manipulated, false, fabricated, or misleading content has a severe negative impact on collective trust, and on the ability of a society to create a common dialogue based on shared accurate facts. It undermines the value of objective facts, delegitimizes experts' voices and authoritative

institutions, and radicalises confrontational stances in a context of chaos and confusion.[1] 'Fake news' would be then a shorthand for a variety of mis-and disinformation vehicles. The repercussions are alarming, but specific quantifiable damage might not be proved, and identifiable injured persons might not be located.

The above-stressed characteristics of misinformation (and disinformation) vehicles have a very relevant effect on their legal analysis, and a direct impact on the components of the liability machinery.[2] If the damage is diffuse, it will be questionable who is entitled to claim compensation, if anyone at all. If the harm is a devaluation of collective trust, it might be difficult to quantify damages to claim. Damage to the public interest is probably the most feared and destabilising impact of the spread of falsity, but it may be not compensable under the principles of the civil liability regime. If the liability system is based on a notice-based scheme, it might be discussed who is expected to report and allege legitimate interests to act. As a consequence, should a fake-news-combating response be addressed and articulated by a liability-oriented discourse, all these considerations must be taken into account to devise the model.

[1] As **the United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information** alert in their *Joint Declaration On Freedom Of Expression And "Fake News", Disinformation And Propaganda* adopted in Vienna, 3 March 2017, (available at , https://www.osce.org/fom/302796 last visit 10/01/2018) (hereinafter, Joint Declaration):

"Taking note of the growing prevalence of disinformation (sometimes referred to as "false" or "fake news") and propaganda in legacy and social media, fueled by both States and non-State actors, and the various harms to which they may be a contributing factor or primary cause;

Expressing concern that disinformation and propaganda are often designed and implemented so as to mislead a population, as well as to interfere with the public's right to know and the right of individuals to seek and receive, as well as to impart, information and ideas of all kinds, regardless of frontiers, protected under international legal guarantees of the rights to freedom of expression and to hold opinions; (...)".

[2] Teresa Rodríguez de las Heras Ballell & Jorge Feliu Rey. *Digital Intermediary Liability or Greater Responsibility: A Remedy for Fake News?* in Weaver, Russell L.; *et al.* (Eds.), *Twenty-First Century Remedies,* North Carolina: Carolina Academic Press, 2019, pp. 91-114. See also: Teresa Rodríguez de las Heras Ballell. *Credibility-enhancing regulatory models to counter fake news: risks of a non-harmonized intermediary liability paradigm shift*. 8(2) J. Int'l Media & Entertainment Law, 2019, pp. 129-162.

On information operations (disinformation) and their potential harm: the role of international law – Talita Dias

Information operations (disinformation) and activities can be defined as any coordinated or individual deployment of digital resources for cognitive purposes to change or reinforce attitudes or behaviours of the targeted audience.[3] This includes a range of online activities, from "harmless" advertising to state-orchestrated fake news campaigns. Such operations have existed for centuries; however, they have garnered increasing attention over the last decade, as states and other stakeholders came to recognise the extent to which digital technologies facilitate their formation and execution and can easily amplify their impact.

Successful information operations do not necessarily coerce targets or wear them down. They influence, persuade, or convince members of the targeted audience to willingly adopt the aims that the author wishes them to adopt, whether by open or covert means.4

The risk of serious harm to states, individuals and other stakeholders is evident, given the range of potential cognitive impacts that information operations can generate. For instance, information operations may destabilise electoral outcomes (e.g., the right-wing occupation of the United States Capitol on 6 January 2021),[5] or they may undermine public health (e.g., the "infodemic" that has disrupted the 'coordinated, medically sound response that is necessary to control the spread of the [COVID-19] virus'[6]). Likewise, they may incite discrimination, violence, genocide, and other atrocities. A paradigmatic example was the dissemination, beginning in 2017, of inaccurate and hateful rhetoric

[3] See Duncan B. Hollis, The Influence of War, The War for Influence, 32 Temple Int'l & Comp. L. J. 30 (2018); Oxford Institute for Ethics, Law and Armed Conflict (ELAC), The Oxford Statement on International Law Protections in Cyberspace: The Regulation of Information Operations and Activities, https://elac.web.ox.ac.uk/the-oxford-Statement-on-the-regulation-of-information-operations-and-activities.

[4] Herbert Lin & Jackie Kerr, *On Cyber-Enabled Information/Influence Warfare and Manipulation*, *in* P. Cornish (ed.), The Oxford Handbook on Cybersecurity (OUP, 2022).
[5] See, e.g., Stuart A. Thompson, Election Falsehoods Surged on Podcasts Before Capitol Riots, Researchers Find, New York Times, 4 January 2022.
[6] Marko Milanovic and Michael N. Schmitt, *Cyber Attacks and Cyber (Mis)information Operations during a Pandemic*, 11 J. Nat'l Sec. Law & Policy 247, 249 (2020).

on Facebook against the Rohingya in Myanmar.[7] More recently, a range of information operations have been employed by the Kremlin to justify the invasion of Ukraine.[8]

Such information operations may fall under different international legal regimes, depending on their author, target audience, and the harm or risk of harm they give rise to. While international law has few specific rules addressing information operations, including dis- and misinformation, it provides a robust and comprehensive legal framework covering different aspects of the phenomenon. Granted, international law only binds states. However, a growing number of domestic legislatures are taking steps to adopt legal frameworks for online harms that mirror international legal obligations, to a greater or lesser extent, and bind online platforms and individuals. Likewise, companies and individuals should abide by international legal standards as a matter of good policy. Navigating the different international legal regimes applicable to the phenomenon of information operations is no easy exercise but acknowledging their existence and striving to clarify how they apply in the digital environment is a good first step.

[7] UN Human Rights Council, Report of the independent international fact-finding mission on Myanmar, Advance Edited Version, UN Doc. A/HRC/39/64, 12 September 2018), paras 73-74 and 84-89; In Myanmar, "pervasive hate speech and shrinking freedom", AL JAZEERA (March 5, 2019).

[8] See BBC Reality Check Team, 'Ukraine crisis: Vladimir Putin address fact-checked', *BBC News*, 22 February 2022, https://www.bbc.co.uk/news/60477712.

II. Rights at Stake: Reconciling Rights and Freedoms – Teresa Rodríguez de las Heras Ballell

Countering misinformation is an extremely delicate process, as it involves a **conflict of rights and freedoms that must be balanced**. Ensuring that the right balance is struck constitutes a fundamental decision in a democratic society.

A number of factors render any attempt to mitigate misinformation very challenging:

- An excessive reaction regarding the removal of misinformation may seriously encroach upon free speech, freedom of information or net neutrality.
- The assessment of falsity/veracity is challenging and multifactorial (context, intent, relevance, knowledge, etc.).
- Unlike illegal content, false content can easily be expressed as mere opinions.
- As mis- and disinformation go viral, attempts at rectifying and correcting them become unsuccessful.

Given the previous analysis, it can be sustained that in the fight against mis/disinformation three categories of content are faced: illegal content, harmful content, and false content. In certain cases, these categories can coincide. The distinction may be relevant for the application of liability rules, and in striking the proper balance between conflicting rights, such as free speech, the right to information, the freedom to run economic activities, and the right to honour. To that purpose, it may be relevant to treat these three categories as distinct and separate ones. Illegal content and false content might not produce actual damage, whereas harmful content could be entirely accurate and truthful, and might be fully licit and legitimate. Therefore, illegality, harmfulness, and falsity constitute different factual spheres that require tailored responses. Hence, preventive measures, and reparation and compensation mechanisms, contrived to combat the effects of illegality and harmfulness, are not equally effective in countering falsity. False content adds intricacies in the detection and assessment phase, and in the ascertainment of damages. The uncontrollable spread of misinformation and disinformation, its penetrating impact on society's stability, and the devastating effects on trust have

crudely revealed such a gap- that is, the lack of preventive and protective measures against falsity. However, in the context of digital platforms' policies and approach to fighting misinformation, the triple distinction may be unnecessary, as digital platforms apply two relevant measures: removal, or demotion. To that end, if an opinion contains verifiable data that is assessed by an independent fact-checker as false, but does not violate the digital platform's policies, it will be demoted but not removed.

Yet, unlike illegal and harmful content, in the case of false content setting a fair balance of the conflicting rights and interests at stake is more complex and unstable. As the contours of false content are blurred, and the potential harm is, whilst albeit severe and massive, highly diffuse, freedom of expression becomes especially vulnerable to any ill-advised restrictive or banning decision.[9]

The gravity of the problem created by coordinated disinformation is not only caused by the falsity of the content or the inauthenticity of the actor, but principally exacerbated by the deafening "noise" its incontrollable penetration and its pervasive expansion produces, silencing authoritative voices and concealing fact-checked content. The risk of 'fake news' is when it becomes widely credible. Factors other than the veracity of facts are able to generate a perception of credibility. Misallocated trust might have a more negative effect than distrust.[10] Fact checking is frequently ineffective in attenuating this wrong perception of credibility, as content is infused with other credibility indicia based on popularity, perception of reliability, or sentiment-driven assessment. Compared to the

[9] As the Joint Declaration states:

1.- General Principles:

a). States may only impose restrictions on the right to freedom of expression in accordance with the test for such restrictions under international law, namely that they be provided for by law, serve one of the legitimate interests recognised under international law, and be necessary and proportionate to protect that interest (...)

2.- Standards on Disinformation and Propaganda

a). General prohibitions on the dissemination of information based on vague and ambiguous ideas, including "false news" or "non-objective information", are incompatible with international standards for restrictions on freedom of expression (...)

[10] Russell Hardin. *Distrust*, 81 B. U. L. Rev. 2001, pp. 495.

widely shared misinformation, the fact-checking and public rectification response might not gain sufficient relevance, and even, when perceived as a minority opinion, it can dilute its credence. Against this backdrop, actions by widely used digital platforms to directly connect people with official and authoritative information, in collaboration with national and global governmental structures, is certainly an avenue to reinforce. During the pandemic, a number of digital platforms and messaging services established Covid-Information centres in collaboration with national ministries of health, as well as the World Health Organisation, to directly reach out to people with accurate health information.

III.- The Role of the Actors Involved in the Fight Against Mis/Disinformation: Digital Platforms Approach - Guillermo Serrano

Mis/disinformation is a highly complex phenomenon that requires a high level of collaboration between the different actors involved (namely digital platforms, but also governments, journalists, independent fact-checking organizations, academia, civil society, and even individual citizens) to be addressed effectively.

As noted in the introduction, the starting point of any strategy to fight misinformation (false information that is often shared unintentionally) and disinformation (coordinated efforts that aim to manipulate or corrupt the public debate seeking a strategic goal, characterised by inauthenticity and coordination) is precisely to clearly distinguish between the two. This distinction is important not only because the policy concerns underlying each differ, but also because the most appropriate response in each case will also differ.

A fundamental distinction in the way we tackle misinformation and disinformation (information/influence operations) at Meta is that we differentiate between the two on the basis of actor/behaviour and content. When we look at disinformation, we focus on actors and their behaviour, while with misinformation the focus is on content. In fact, actors engaged in information/influence operations (disinformation) need not necessarily use false information, and it would in fact be acceptable political discourse if it was shared by authentic actors. The real issue is that the actors behind these campaigns are using deceptive behaviours to either conceal the identity of the

organisation behind a campaign, make the organisation or its activity appear more popular or trustworthy than it is, or evade enforcement efforts.

1. Misinformation

Given the difficulties, dilemmas and the potential clashes between fundamental rights involved in addressing misinformation (those of free speech, the right to receive truthful information, the right to safety and security etc.), our overarching objective at Meta is to strike the right balance between having an authentic and secure platform whilst respecting fundamental rights like free speech. Furthermore, we consider that it is neither possible, nor desirable, that a private company acts as a sort of 'arbiter of truth'. We simply lack the legitimacy to do so. Hence, we rely on our collaboration with independent fact-checking organisations (more details about how this collaboration works in practice are to be found below).

Our approach to misinformation applies a **three-part strategy - remove, reduce, and inform** - to manage problematic content across the Meta family of apps (Facebook and Instagram).

(1) We **remove** content that violates our policies (Community Standards), including (but not limited to): fake accounts and accounts engaged in inauthentic behaviour, misinformation that may contribute to the risk of imminent violence or harm (for instance, COVID-19 misinformation claims debunked by the WHO that could cause harm in the real world, like 'drinking bleach cures COVID-19'), and voter fraud or interference (which includes any misrepresentation about how to participate in the voting process, such as the dates, location, time, methods, and qualification).

(2) We reduce the distribution and visibility of problematic content that does not violate our policies, but still undermines the authenticity of the platform. For instance, content debunked by independent third-party fact-checkers is demoted in the Feed. In demoting content instead of removing it we aim at striking a balance between having an authentic and secure platform whilst respecting fundamental rights like free speech.

For this important task, we rely on our collaboration with independent fact-checkers. We have a global partnership with the International Factchecking Network (IFCN) and collaborate with IFCN certified fact-checker's partners at the country level. Facebook's approach works by identifying content to be

reviewed by independent fact-checker's partners through a combination of Technology (ML), user complaints/reports, human review, and fact-checkers themselves. Fact-checkers then choose which content to review and rate. Finally, Meta acts on the content rated as false by fact-checkers by demoting it in the Feed. In addition, Pages that repeatedly share content that is rated false by fact-checkers will see their Page distribution reduced in the News Feed, and their ability to monetise and advertise removed. We have expanded this policy to also include penalties for individual accounts.

(3) Finally, we **inform** people with **additional information and context** (for example, by adding misinformation labels to content across Facebook and Instagram that has been rated false or partly false by independent fact-checkers, or adding a warning message for people trying to share content labelled as misinformation – amongst other actions) **and connect people with authoritative and/or official sources of information so they can make informed decisions** (by way of example, during the pandemic we launched our COVID-19 Information Centre at Facebook and Instagram, with real-time updates from national health authorities and global organizations, such as the World Health Organization).

2. Disinformation (information/influence operations)

When it comes to disinformation, the strategy and concerns are different (as noted above). If the main challenge when dealing with misinformation is to strike the right balance between having an authentic and secure platform whilst respecting fundamental rights like free speech, in the case of disinformation we do not tolerate any such activity and take these actors and their content down as soon as we become aware of this behaviour.

The most egregious form of this type of deception is Coordinated Inauthentic Behaviour (CIB): that is, any coordinated network of accounts, Pages and Groups on our platforms that centrally relies on fake accounts to mislead Meta and the people using our services about who is behind the operation and what they are doing. There are two types of CIB: (a) Foreign-led efforts to manipulate the public debate in another country, and (b) Operations run by domestic, state and/or non-government actors.

In order to find and act against influence operations, we focus on behaviour, as that is the best way to stop the abuse. Hence, our investigative work and enforcement are location, and are content agnostic. They actively look for the elements common to every information operation: (1) Coordination among

accounts, among Pages, or among offline groups; (2) Manipulation or deception; and (3) A strategic goal to influence public discourse.

With each investigation, we identify the behaviours that are common across the people trying to do harm. Then, we work to automate the detection of these behaviours, and even modify our products to make those behaviours much more difficult. If expert investigations are like looking for a needle in a haystack, our automated work is like shrinking that haystack.

IV. Rules on Platforms and Social Media: Enhanced Responsibility and Liability (From 'Safe Harbour' Systems Onwards) - Teresa Rodríguez de las Heras Ballell

States can decide to devise a legal framework that, without regulating or interfering in the information sector and the activity of social media, does however effectively allocate incentives for private actors. That is one of the traditional goals of civil liability rules. Along with the goal of compensating victims, civil liability rules aim to deter those activities that a particular society perceives as undesirable or inacceptable, and to encourage private actors to adopt precautionary measures to prevent such harmful activities. To that end, liability risks should be allocated on the "cheapest-cost avoider". Rules on enhanced responsibility recently developed in the European Union (the *Recommendation on tackling illegal content*), and the key 'safe harbour' regime (*E-Commerce Directive*[11] *and the Digital Services Act*[12]) are very important incentive-allocating policy decisions that will define the contributory role of platforms in the fight against disinformation.

The core debate has been to what extent a liability-based approach in the digital services and digital platform legal framework would be an advisable solution, and how much it would contribute to striking the balance between conflicting rights. From that perspective, intermediaries and digital platforms represent a critical component in the dis-misinformation machinery. It is undeniable that

[11] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce').

[12] Regulation of the European Parliament and the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.

intermediaries and digital platforms provide the infrastructure that bad actors exploit for the dissemination of their content, create an environment suited to ignite the perception of credibility, and exacerbate the massive effects of false news. Nevertheless, it is highly questionable that such an infrastructural contribution should hold [any level of liability. More interestingly, it is even more uncertain how digital platforms should act to contain virality, counter popularity-measured credibility, and combat with objectivity and fact checking an oversized perception of trustfulness. A regulatory model that happens to dislocate incentives may trigger an overly cautious reaction of intermediaries and digital platforms, for fear of the liability consequences, likely to distort the free flow of ideas in the digital world, and to encroach upon freedom of expression.[13]

Therefore, a liability-based approach to digital platforms is neither advisable nor effective. A liability solution is not going to replace the "lack of legitimacy" of digital platforms as private companies to arbitrate freedom of expression and other fundamental rights. Thus, a responsibility approach seems instead to better promote cooperation among different actors. Overall, social media and content-sharing platforms have already implemented proactive and very effective programs to prevent, react, and mitigate misinformation, such as notice-and-take down systems, trusted flaggers, internal policies, account blocking, fact-checkers, information centres in collaboration with public institutions, and so forth.

[13] Teresa Rodríguez de las Heras Ballell. Credibility-enhancing regulatory models to counter fake news: risks of a non-harmonized intermediary liability paradigm shift. 8(2) J. Int'l Media & Entertainment Law, pp. 129-162, 2019.

PROPOSALS AND RECOMMENDATIONS

A) THE ROLE OF PLATFORMS - Should platforms be liable or otherwise responsible for user-generated dis/misinformation? What is the appropriate legal or regulatory approach to address the 'infodemic' and related influence or information operations, such as online hate speech? Should states aim for a liability, responsibility/accountability, or duty of care model?

The Working Group (WG) concludes that the intermediary liability model is inappropriate for tackling harmful content, both illegal and legal, because it will inevitably force online platforms to err on the side of censorship, i.e., it will lead to the over-removal of content by platforms to avoid liability and/or the accompanying prohibitively high fines. Furthermore, the WG agrees that, from a private law perspective, liability is often associated with causing direct damage to a specified individual, which is at odds with the type of diffuse or collective damage generated by dis/misinformation and other information/influence operations in society. Instead, **a model focussing on the responsibility of online platforms for failing to exercise, for instance, a duty of care, i.e., their best efforts to assess the risks and mitigate the presence of illegal, false, or harmful content, would strike a better balance between freedom to receive and impart information and the protection of individuals and society from online harms.**

In human rights law terminology, fighting dis/misinformation and other online harms with strict intermediary liability rules and policies may not only be unnecessary to achieve legitimate aims (e.g., the protection of health and public order), but may also impose a disproportionate burden on the right of individuals to freedom of thought, information, expression, and participation in democratic processes, among others. Likewise, the **WG recommends that, whatever regulatory model is chosen in whichever political system, any official oversight body must be sufficiently independent of the executive, legislative, and judiciary branches**. If formally part of any such branch, states should avoid granting excessive power to official oversight bodies. They should also ensure that powers belonging to other government branches are not usurped by any such body. This is necessary to ensure a fair separation and equitable power balance between the various branches of government over the online information environment. It is also imperative for a free and pluralistic media environment, offline and online.

Likewise, given the transboundary nature of online harms, the WG recommends that states consider how to cooperate with other states when regulating online platforms, such as by agreeing on specific international digital media standards or setting up international media councils with an advisory function.

Finally, the WG believes that, before enacting legislation or regulation in this area, multiple stakeholders are consulted, including online platforms themselves, academia, and civil society organisations.

B) CONTENT MODERATION - To what extent should states and online platforms separate their approach to incitement to violence, hostility, or discrimination from how they tackle other types of problematic content? How difficult is it for online platforms to make such content moderation decisions?

The WG members have carefully considered this question and agree that States should clearly distinguish, by law, between content that gives rise to an imminent risk of violence, such as the tweets that instigated the January 6th Capitol riots, and other, less serious speech acts, such as misinformation about the origins of the COVID-19 pandemic. This type of framework already exists under international and domestic human rights law as well as in several constitutional systems around the world.

For instance, under Article 20 of the International Covenant on Civil and Political Rights (ICCPR), states must prohibit by law any war propaganda, as well as any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility or violence. Breach of such prohibition need *not* give rise to criminal liability of the individual speaker, but may amount to civil liability, depending on the seriousness of the speech act in question. Relevant factors include the speech act's proximity to the relevant outcome, the speaker's intention, and the severity of the content itself.

On the other hand, while states are *not* in principle required to prohibit other types of online speech, they *may* limit the freedom to receive and impart information to protect a legitimate aim, in line with Article 19(3) of the ICCPR. In some circumstances, the protection of other fundamental human rights at risk, such as the rights to health, life, and non-discrimination, may even require states to limit individuals' right to receive and impart information. However, states may only limit free speech in accordance with the law, for a legitimate purpose, in a necessary and proportionate manner. The

same requirements (legality, legitimacy, necessity, and proportionality) feature in Article 10(2) of the European Convention on Human Rights, now part and parcel of the domestic legal system of several European countries. These requirements should apply not only to the definitions of limited speech but also to their respective limiting measures.

This all means that, when moderating content on a daily basis, online platforms should follow human-rights consistent laws and regulations. In the absence of such laws and hard-pressed to make difficult calls about content posted on their platforms, companies should strive to look to international human rights law for guidance, in line with the United Nations Guidelines on Business and Human Rights.

The WG recognises that online platforms are under constant pressure from a variety of social and political groups – including opposing voices – to act upon content that such groups perceive as false or harmful. If platforms decide to leave 'problematic' content up, they are accused of endorsing it and so are pressured to remove it. On the other hand, if platforms actively moderate content that goes against their community standards but is perceived as unproblematic or legitimate by certain members of the public, they are accused of clamping down on free speech. The WG also acknowledges the difficulty of moderating an unsurmountable number of posts at scale every single day. Moderating dis/misinformation may be particularly difficult given the nuances and contextual nature of language, as well as the fine line that often exists between fact, fiction, and opinion, especially in cases where the science is not yet fully settled. The same is true of other types of problematic content, such as online hate speech, for which contextual interpretation is key to assess the very type of content at stake.

In addition, online platforms often lack the technology to make such moderation decisions at scale in a speedy manner. For instance, while artificial intelligence technology is effective at identifying terrorist, child abuse, and pornographic content, especially images, the same cannot be said of other types of content, for which moderation still requires human judgement.

In short, moderating content, especially on large online platforms, is no easy task, and policymakers and civil society should acknowledge that.

C) TOOLS TO TACKLE DIS/MISINFORMATION - What are the current and potential tools at the disposal of online platforms to tackle dis/misinformation and other types of problematic content?

The WG notes that several tools are available to online platforms when tackling or moderating dis/misinformation and other types of problematic content. These include measures that have already been put in place in the context of COVID-19 dis/misinformation, such as labelling or tagging a certain item of content as potentially misleading, dismantling bots, false personas, or orchestrated disinformation campaigns, deprioritising/demoting such content or taking it down in extreme cases, as well as directing users to content produced by official health bodies, such as the World Health Organization. **The WG welcomes these measures.** However, **it also recommends that other tools to tackle dis/misinformation in a human rights-compliant manner be developed, tested, and put in place by online platforms, including in cooperation with other relevant stakeholders, such as states, academia and civil society organisations**.

These additional measures include platform-driven media and information literacy campaigns, as well as greater transparency in content moderation decision-making processes, such as by clearly informing users about content flagging and takedown processes and providing comprehensive data on the decisions or outcomes of those processes. **The WG also recommends that the adoption of these various measures be calibrated to the type of content being moderated (e.g., illegal, harmful, or false; disinformation vs misinformation), as well as its potential risks and impact on societal interests and individual rights (e.g., health, life, and electoral integrity).**

The WG notes that the phenomenon of dis/misinformation has had an unprecedented effect on societal stability and unity, as well as on trust in public institutions. Given the ease with which false, hateful and sensationalist content can be disseminated online, coupled with the pervasiveness of online platforms in societies across the globe, dis/misinformation often becomes viral and uncontrollable.

The WG also recognises the role that platform recommendation algorithms may play in the dissemination and virality of dis/misinformation and other types of problematic content.

This is because such algorithms may be designed to maximise user engagement and thereby generate online advertisement revenue, with the inadvertent effect of fuelling problematic content.

The WG agrees that, despite their ongoing efforts to tackle dis/misinformation and other types of problematic content, online platforms can and should do more to address the issue by preventing the exploitation and inadvertent impact of recommendation algorithms.

- One option on the table is to give users more control over what they see on their feeds or search results. This could be done, for instance, by allowing users to directly adjust their personal content curation parameters or by developing advertisement-free versions of their platforms for a fee.

- Another set of options is to increase transparency around how content is curated and thus made visible to each user based on their personal data gathered online. This could include enabling users to understand why each piece of content has reached them, as well as disclaiming or labelling sponsored content or political advertisement (of note, both of these measures have already been put in place by some online platforms).

- Other options that the WG recommends, which remain underexplored, include:

a) third-party auditing or verification of recommendation algorithms, done consistently with platform proprietary rights, as well as

b) the publication of the data used to train and design such machine-learning algorithms, along with accuracy indicators and other test results.

D) THE ADVERTISEMENT-BASED BUSINESS MODEL - Is there a feasible alternative to online platforms' advertisement business models? For instance, how can states and civil society ensure that online platforms act with greater public interest?

The WG is mindful of the power and impact of online platforms in societies around the world, including through their recommendation algorithms and advertisement business models. At the same time, **WG members also recognise that it may be financially unsustainable for companies to do away entirely with their advertisement business models**. Likewise, it would be difficult, if not impossible, to imagine how less advantaged or marginalised societal groups could continue to have access to online platforms and their rights-enhancing potential if such platforms followed a pure advertisement-free, paid subscription-only model. **The WG thus agrees that the key lies in striking**

the right balance between on the one hand enabling platforms to operate in a profitable environment, and thereby remaining accessible to the majority of the population, and on the other reducing their human rights impact and other adverse consequences.

There is no single, one-size-fits-all answer to this conundrum. This is first and foremost due to societal, cultural, political, and economic differences across states. Whilst, in some countries, significant parts of the population could afford to pay for online platform access or the state itself could subsidise such access, in others, this is not an economically feasible option. Moreover, it is also imperative to ensure that both traditional and online media outlets remain independent from the state. Thus, states should carefully consider whether and to what extent they should influence or affect individuals' access to online platforms.

Nevertheless, the WG believes that imbuing public interest values or elements within private technology companies is a feasible option that may counteract some of the negative human and societal impact of platforms' business models.

The WG also recommends that states have in place robust competition laws, and other ex-ante regulatory frameworks (e.g., the Digital Markets Act in the EU) to ensure that market access among small and large platforms is shared in a fair, equitable, and transparent manner.

E) DEMOCRACY, SCIENCE, AND DIS/MISINFORMATION - To what extent has dis/misinformation undermined the authoritative nature of science?

A significant number of dis/misinformation campaigns on social media and other online platforms are aimed at discrediting well-established scientific knowledge. The paradigmatic case has been COVID-19-related dis/misinformation. Examples range from false or misleading content offering ineffective or harmful alternative cures to posts discrediting the efficacy or safety of approved vaccines. Other examples include health dis/misinformation more generally, and false claims about the man-made origins of climate change and its devastating consequences. Thus, **the WG concludes that dis- and misinformation have, to some extent, undermined trust in science and its processes, especially among less-educated or resilient groups**. While dis- and misinformation are not exactly new phenomena, in the digital age, they have reached an unprecedented virality given the scale, pervasiveness and speed of the Internet.

Despite the importance of the freedom to receive and impart information of all kinds in democratic societies, states and other stakeholders broadly agree that such freedom is not absolute. For example, Article 19(3) of the ICCPR, currently ratified by 173 states around the world, provides that the exercise of those rights 'carries with it special duties and responsibilities' and may thus be 'subject to certain restrictions. These restrictions are valid insofar as they are 'provided by law and are necessary' to respect the rights or reputations of others and for the protection of national security or public order, public health, or morals.

Public health features prominently among one of the rights or values that ought to be balanced against free speech and should often prevail over it. Likewise, the right to *life* is a core human right whose protection may often justify limitations to the freedom to receive and impart information. Therefore, **the WG agrees that health dis/misinformation may be limited insofar as it threatens the life and health of individuals**. However, both groups recognise that this approach requires caution and a careful balancing between conflicting human rights. Thus, **it is recommended that only in cases where dis/misinformation poses a *significant and imminent* risk to life or health, such types of false or misleading content should be taken down. When no such risk is present, other, less serious limitations, such as labelling or content redirection, should be adopted**.

It is often the case that politicians – both in the executive and legislative branches – are themselves the sources or disseminators of false or misleading information about COVID-19; for instance, by promoting unproven, ineffective, or harmful treatments, or by discrediting scientifically proven ones, such as clinically-approved vaccines. Such statements may have a particularly devastating impact on a state's public health response to the pandemic, given the prominence of the speakers involved. At the same time, when the science is uncertain (as was the case early on in the pandemic when it was unclear how the virus was transmitted), fair and democratic political discourse on public health policy must not be undermined. Thus, **the WG agrees that, if the content of the political discourse is of such a character as to pose a significant and imminent risk to individual life or public health, it should not deserve any special curation or moderation privileges**. Quite the contrary: such speech acts should be treated like any type of serious health dis/misinformation, if not more stringently, given the risk that content produced or endorsed by such actors may disseminate and convince others more easily. Thus, **some members of the WG propose the adoption of Codes of Conduct for political speakers to guide their behaviour in the online and offline information environment**. At the same time, other, less serious, types of political health dis/misinformation should be tackled with greater care, balancing the need for free, democratic, and pluralistic political discourse with the protection of life, health, and other important societal values.

F) DATA SOVEREIGNTY - How to tackle dis/misinformation in states that follow a 'data sovereignty' model, i.e., whereby states have control over the data and information made available to the public?

The idea that states should have sovereignty over the data or information that may be accessed by the public is a euphemism for state censorship, that is, state-imposed limitations to freedom to receive or impart information and democracy more generally.

Thus, in those states, it is even more challenging to fact-check and counter state-endorsed narratives, including those amounting to dis/misinformation. In fact, in many such states, the media—both traditional and online—is controlled by the state, and foreign online platforms such as Google or Facebook are banned from operating therein. The WG agrees that there is no easy way to tackle this problem, especially because Western democracies often see it as a distant threat to the human rights of foreigners. However, in a world where the internet and pandemics do not follow territorial boundaries, states and platforms must work together to counter dis/misinformation in non-

democratic countries too. **The WG believes that education, especially media and information literacy, is an essential tool to tackle the problem at this root**. This may require, for example, greater investment in cultural or educational exchanges between Western and Eastern societies, as well as Global North and South communities.

G) EDUCATION AND ONLINE RESILIENCE - Does the root of the 'infodemic' and other 'information disorders' lie in the public's lack of sufficient education and/or online resilience?

There was consensus among WG members that, at their core, dis/misinformation campaigns during the COVID-19 pandemic, and other influence or information operations, have been driven and made successful by the lack of public awareness and education about the online information environment and how it operates. As all kinds of information have become more easily, and often overwhelmingly, available online (at the touch of one button, a click, or a scroll), individuals have spent less time and effort educating themselves about the sources and verifiability of such information. Likewise, little attention has been paid by users to opposing narratives and, crucially, to the means by which such information is made available to them (i.e., the content curation or recommendation process). The digital age is an age of convenience, which is not conducive to careful research, thinking and speaking.

The WG agrees that the best way to address this issue is, once again, by increasing education levels among the public, particularly by promoting media and information literacy campaigns or initiatives.

Media and information literacy includes the ability to understand a) the information or content itself that is presented; b) how this content has been generated and made accessible to an individual; and c) how this individual can respond to it. As noted by the United Nations General Assembly in its March 2021 Resolution, media and information literacy can help build resilience in societies, ensuring that online and offline media users are less vulnerable to false or misleading content. This should in turn prevent, or at least mitigate, the further spread and damaging effects of dis/misinformation in societies. To achieve that, **the WGs**[14] **agree that both states and private actors must play a crucial**

[14] See paper on Nationalism in the context of the COVID-19 pandemic

role in developing and fostering global, regional, and local media and information literacy initiatives, including by cooperating with one another. They also recognise that existing economic, digital, gender and intersectional divides must be reduced, especially between developed and developing countries, as well as between privileged and marginalised groups within each country.

Finally, the WG agrees that a free, pluralistic, and robust media environment, online and offline, is essential to ensuring the success of media and information literacy initiatives.

H) WELFARE MODELS - Is a social or welfare state model the necessary 'middle-ground' to build public resilience, increase the level of education in societies, fight inequality and curb the current information disorder?

The WG discussed holistic political options to tackle nationalism, health dis/misinformation and the information disorder at their core. It was pointed out that especially middle- and lower-income classes in Western democracies have been wary of, or disengaged with, the idea of a welfare state. This is perhaps due to an 'anti-establishment' sentiment, grown out of disappointment with traditional social-liberal policies, coupled with the rise and spread of populist rhetoric, including via online dis/misinformation.

The WGs[15] conclude that a social, welfare state remains the best political option available to balance the need to invest in strong public health and educational policies, on the one hand, and the importance of free, open, and democratic information spaces on the other.

Thus, both the WGs recommend a continued promotion or revival of a welfare state in democracies around the world, including by debunking the myths that have fuelled an unwarranted anti-establishment sentiment.

[15] See footnote above (186)

AUTHORS:

- **Dr Talita Dias,** Shaw Foundation Junior Research Fellow in Law, Jesus College, Oxford.
- **Teresa Rodríguez de las Heras Ballell,** Professor of Commercial Law at the Carlos III University of Madrid. Sir Roy Goode Scholar at UNIDROIT. Academic Advisor to Fide. *(Leader of the Group)*
- **Guillermo Serrano,** Public Policy Manager, Spain & Portugal, Meta. Currently Head of Government Relations in Trainline España

NATIONALISM, POPULISM, AND IDENTITIES:

CONTEMPORARY CHALLENGES

JANUARY **2023**

FIDE FOUNDATION

THINKFIDE.COM