



The Hague Centre  
for Strategic Studies

# Red Lines & Baselines

Towards a European Multistakeholder  
Approach to Counter Disinformation

**Authors: Louk Faesen, Alexander Klimburg, Simon van Hoeve, Tim Sweijs**

**Introductions by Dennis Broeders, Alexander Klimburg, Frederick Douzet, Chris Marsden and Trisha Meyer**

October 2021



The Hague Centre  
for Strategic Studies

## Red Lines & Baselines

Towards a European Multistakeholder  
Approach to Counter Disinformation

### Authors:

Louk Faesen, Alexander Klimburg, Simon van Hoeve, Tim Sweijts

### Introductions by

Dennis Broeders, Alexander Klimburg, Frederick Douzet,  
Chris Marsden and Trisha Meyer

ISBN/EAN: 9789492102836

October 2021

© *The Hague* Centre for Strategic Studies. All rights reserved.

No part of this report may be reproduced and/or published in any form by print, photo print, microfilm or any other means without prior written permission from HCSS. All images are subject to the licenses of their respective owners.

The research for and production of this report has been conducted within the PROGRESS research framework agreement. Responsibility for the contents and for the opinions expressed, rests solely with the authors and does not constitute, nor should be construed as, an endorsement by the Netherlands Ministries of Foreign Affairs and Defense.

The Hague Centre for Strategic Studies

info@hcss.nl

hcss.nl

Lange Voorhout 1

2514EA The Hague

The Netherlands

Cover adapted from 7CO's "Kundgebung (3) Protest against corona measures" and licensed under CS BY 2.0. Some rights reserved.

# Contents

<b>List of Abbreviations</b>	<b>IV</b>
<b>Executive Summary</b>	<b>V</b>
1. Proceed carefully with a government-to-government norm	VI
2. Agree on an industry charter of standards for social media platforms	VI
3. Develop a coregulation model to advance the standards from formulation to implementation	VII
4. Establish a Disinformation Information Sharing and Analysis Center (DISINFO-ISAC)	VIII
<b>1 Introduction</b>	<b>1</b>
<b>2 Introduction to disinformation: definitions and the regime complex</b>	<b>4</b>
2.1 Definitions	4
2.2 The Regime Complex for Countering Disinformation	6
2.3 Key Takeaways	8
<b>3 Norms</b>	<b>9</b>
3.1 Introduction by Dennis Broeders	9
3.2 Norms Primer	11
3.1 Big N Norms	13
3.2 Small n norms	15
3.3 Key Takeaways	16
<b>4 Big N Norm against disinformation set by governments 17</b>	
4.1 Introduction by Alexander Klimburg	17
4.2 The basis for a norm against disinformation	19
4.3 Framing and linking a norm against disinformation	22

4.4	Promoting the norm against disinformation	23
4.5	The risks of a disinformation Norm	23
4.6	Key Takeaways	28
<b>5</b>	<b>Small n norms against disinformation set by the industry</b>	<b>29</b>
5.1	Introduction by Frederick Douzet	29
5.2	Small n norms Primer	31
5.3	Moderation through the Tech Stack	32
5.4	Small n norms for Social Media Platforms	36
5.4.1	Community Guidelines	37
5.4.2	Bot Takedowns	38
5.4.3	Factchecking	41
5.4.4	Labelling	42
5.4.5	Political Advertising	44
5.4.6	Verified Information Features	47
5.4.7	Algorithms and automated moderation	49
5.4.8	Community Reporting and Remediation	52
5.5	Key Takeaways	54
<b>6</b>	<b>A coregulation model to advance the standards</b>	<b>56</b>
6.1	Introduction by Chris Marsden and Trisha Meyer	56
6.2	Coregulation Primer	59
6.3	The Regulatory Regimes	59
6.4	Why Coregulation?	61
6.5	What a proposed coregulation model could look like	62
6.6	Institutionalizing coregulation: A short proposal for a DISIN-FO-ISAC	66
6.7	Key Takeaways	72
<b>7</b>	<b>Conclusion and Recommendations</b>	<b>73</b>
<b>8</b>	<b>Annex I: industry best practices for countering disinformation</b>	<b>77</b>
8.1	Labelling	78
8.2	Community or Voluntary Reporting	85

8.3	Third-Party Factcheckers	90
8.4	Oversight Boards	94
8.5	Community Guidelines	97
8.6	Algorithmic and automated content moderation	102
8.7	Verified Information Features	108
<b>9</b>	<b>Annex II: Platform Governance and Cooperation across the Regime Complex: Case Insights</b>	<b>111</b>
9.1	Case Study 1: The Global Internet Forum to Counter Terrorism (GIFCT)	112
9.1.1	Background	112
9.1.2	The GIFCT	113
9.1.3	Analysis and Conclusions	115
9.1.4	Recommendations	118
9.2	Case Study 2: The EU Code of Practice on Disinformation	120
9.2.1	Background	120
9.2.2	The Code of Practice	121
9.2.3	Analysis and Conclusions	122
9.2.4	Recommendations	125
9.2.5	The Future of the Code	128
<b>10</b>	<b>Annex III: The Regulatory Regimes</b>	<b>129</b>
10.1	Option 0 – Status Quo.	129
10.2	Option 1 – Self-regulation	130
10.3	Option 2 – Coregulation	131
10.4	Option 3 – Statutory Regulation	132
<b>11</b>	<b>Annex IV: ISACs - the Gold Standard in Cybersecurity Information Exchange</b>	<b>134</b>
<b>12</b>	<b>Annex V: Interviewees</b>	<b>138</b>
	<b>Bibliography</b>	<b>139</b>

# List of Abbreviations

<b>AP</b>	The Associated Press	<b>ICFN</b>	International Fact-Checking Network
<b>API</b>	Application Programming Interface	<b>ICJ</b>	The International Court of Justice
<b>ASPI</b>	Australian Strategic Policy Institute	<b>IETF</b>	The Internet Engineering Task Force
<b>CDN</b>	Content Delivery Network	<b>ISAC</b>	Information Sharing and Analysis Centre
<b>CIGI</b>	The Centre for International Governance Innovation	<b>ISO</b>	International Organization for Standardization
<b>CIP</b>	Content Incident Protocol	<b>ISOC</b>	The Internet Society
<b>CISA</b>	United States Cybersecurity and Infrastructure Security Agency	<b>ISIS</b>	Islamic State of Iraq and Syria
<b>CSAM</b>	Child Sexual Abuse Material	<b>ISP</b>	Internet Service Provider
<b>CSIS</b>	Centre for Strategic and International Studies	<b>ITRs</b>	International Telecommunication Regulations
<b>DDoS</b>	Distributed Denial-of-Service attack	<b>ITU</b>	International Telecommunication Union
<b>DFR Lab</b>	Digital Forensic Research Lab (The Atlantic Council)	<b>IWF</b>	Internet Watch Foundation
<b>DG CNECT</b>	Directorate-General for Communications Networks, Content, and Technology	<b>KPI</b>	Key Performance Indicator
<b>DHS</b>	United States Department of Homeland Security	<b>MANRS</b>	Mutually Agreed Norms for Routing Security
<b>DIGI</b>	Digital Industry Group Inc	<b>NAI</b>	Network Advertising Initiative
<b>DNSSEC</b>	Domain Name System Security Extensions	<b>NetzDG</b>	The Network Enforcement Act (also known as <i>Netzwerkdurchsetzungsgesetz</i> or <i>Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken</i> )
<b>DSA</b>	Digital Services Act	<b>OII</b>	Oxford Internet Institute
<b>EDMO</b>	European Digital Media Observatory	<b>PEGI</b>	Pan European Game Information
<b>EEAS</b>	European External Action Service	<b>RACER</b>	Relevant, Accepted, Credible, Easy to monitor, Robust
<b>EC</b>	European Commission	<b>RAS</b>	Rapid Alert System
<b>ECHR</b>	European Convention on Human Rights	<b>RUSI</b>	Royal United Services Institute
<b>EDRi</b>	European Digital Rights	<b>SAV</b>	Source Address Validation
<b>EEAS</b>	European External Action Service	<b>STIX</b>	Structured Threat Information eXpression
<b>EFF</b>	Electronic Frontier Foundation	<b>SCO</b>	Shanghai Cooperation Organization
<b>ERGA</b>	European Regulators Group for Audiovisual Media Services	<b>TAXII</b>	Trusted Automated eXchange of Intelligence Information
<b>EU</b>	European Union	<b>ToS</b>	Terms of Service
<b>FTO</b>	Foreign Terrorist Organization	<b>UDHR</b>	Universal Declaration of Human Rights
<b>GARM</b>	Global Alliance for Responsible Media	<b>UN</b>	United Nations
<b>GDPR</b>	General Data Protection Regulation	<b>UN GGE</b>	UN Group of Governmental Experts on Developments in the field of information and telecommunications in the context of international security
<b>GIFCT</b>	Global Internet Forum to Counter Terrorism	<b>UN OEWG</b>	Open-ended working group on developments in the field of information and telecommunications in the context of international security
<b>GMF</b>	The German Marshall Fund	<b>US</b>	United States
<b>GNI</b>	Global Network Initiative	<b>UK</b>	United Kingdom
<b>HLEG</b>	The High-Level Expert Group on Fake News and Online Disinformation	<b>WARP</b>	Warning, Advice and Reporting Point
<b>Hybrid CoE</b>	European Centre of Excellence for Countering Hybrid Threats	<b>WCIT</b>	World Conference on International Telecommunications
<b>IARC</b>	International Age Rating Coalition		
<b>ICANN</b>	The Internet Corporation for Assigned Names and Numbers		

# Executive Summary

Disinformation continues to dominate the headlines as one of the major political challenges that liberal democracies face today. In recent years, disinformation campaigns have exacerbated existing political polarizations with effects ranging from discrediting measures against the COVID-19 pandemic to inciting mass violence against the very institutions of democracy itself. They have undermined trust in democratic processes and even had a significant impact on elections and plebiscites across North America and Europe. Because of the intricate relationship between disinformation and civil rights – most notably freedoms of speech and expression – governments, industry and civil society alike still struggle to find effective ways to counter this threat.

Given the complexity of the counter-disinformation ecosystem, consensus on a single legal solution that is widely shared by all actors is highly unlikely. For that reason, so-called norms can be useful tools, not only because of their dynamic and flexible nature, but also because they can be established by other stakeholders than just governments. Norms are voluntary, legally non-binding commitments that reflect a common standard of acceptable and proscribed behavior, accompanying and expanding on existing legal understandings.

This report specifically looks at norm development against disinformation in an international “Whole of System” context examining two approaches: a government-to-government norm (or “big N Norm”) and an industry agreement or charter of standards for social media platforms (or “small n norms”). For the big N Norm approach the report evaluates the feasibility of such a norm in light of highly diverging notions of *information security*. For the small n norms approach, the report considers an industry charter and proposes eight standards for social media platforms, their requirements and risks. Finally, a coregulation model is proposed to advance the industry charter, along with a Disinformation Information Sharing and Analysis Centre (DISINFO-ISAC) to help operationalize this model and facilitate industry cooperation on threat information sharing in the context of disinformation.

The small n norms approach is the favored option as it offers more proportionality in moderating online conduct and content, avoids contentious diplomatic and civil rights issues, such as governments deciding on what constitutes ‘good’ or ‘bad’ content, but neither does it leave such judgments to the sole discretion of companies. The industry charter overlaps with the suggested proposals of the European Commission to enhance the EU Code of Practice on Disinformation offered in May 2021, but offers more concrete language for each standard, as well as their requirements, risks and the key performance indicators to incentivize transparency and monitor the implementation by the social media platforms. The charter, along with other standards proposed elsewhere, can therefore inform the European Democracy Action Plan and the deliberations of the signatories of the Code as they prepare an update by late autumn 2021. The proposed coregulation model, in turn, introduces enforcement and compliance mechanisms.

Overall, the following four recommendations are made to encourage restraint from state actors and improve resilience of industry actors with the aim of countering disinformation. While directed primarily at governments and industry, most recommendations will require a concerted multistakeholder approach.

## 1. Proceed carefully with a government-to-government norm

A big N Norm against disinformation can be framed around covert election interference and linked to the nonintervention principle. After all, election meddling is one of the few forms of disinformation that appears to reach the *coercion* threshold of the nonintervention principle on the basis of official statements or responses from Western like-minded countries. Such a specific norm proposal, however, would risk providing support to pre-existing Russian proposals made in the UN First Committee and elsewhere of ‘total noninterference’. These Russian proposals prioritize sovereignty over universal human rights, and a multilateral Internet governance system over the existing multistakeholder model. It would also risk opening the door to the Russian concept of *information security* and an intergovernmental discussion on what content should be allowed online, allowing human rights abusers to argue for sovereign control of information and to crack down on dissenting voices via censorship.

At the same time, such a big N norm would prohibit concerted foreign disinformation campaigns and *covert* influence operations aimed at undermining democratic processes, while allowing *overt* measures. The suggestion above would form a compromise of sorts: *overt* means of any type, including strategic communication or ‘propaganda’ by state media actors such as RT (or from a Russian point of view BBC or CNN), would be considered acceptable, as would publicly declared funding of civil society organizations (including, for instance, the US National Endowment of Democracy or the Russian Russkiy Mir Foundation). However, covert subterfuge, malinformation, or non-transparent strategic communication would be unacceptable.

While such a norm proposal can very well be a viable option that focuses more on conduct rather than content, thereby alleviating some human rights concerns, it comes with considerable drawbacks that would require special care from a norm entrepreneur to avoid inadvertent interpretations of the norm.

## 2. Agree on an industry charter of standards for social media platforms

Developing “small n norms” is considered a less risky, preferable alternative that is less prone to civil rights abuses. These standards are based on an extensive analysis of the EU Code of Practice on Disinformation (see Annex II) and existing best practices from the social media companies (see Annex I). They should be seen as a minimum standard – not the end-goal but a starting point from which resilience in countering disinformation is built.

Industry standards do not come without any risks for civil rights and the already dominant role of social media platforms as the arbiters of truth. They would require careful implementation associated with transparency and accountability measures. Eight standards are therefore suggested that include requirements for social media platforms (e.g. common definitions), potential risks and key performance indicators (KPIs) (see Table 1). Taken together they can be considered for an industry charter or as part of the upcoming update of the EU Code of Practice on Disinformation. Resources can be directed to a survey with government, industry



and civil society stakeholders to assess the feasibility and effectiveness of the eight standards proposed in this report.

*Table 1 Eight proposed standards for an industry charter.*

Standard	Description
<b>Community Guidelines</b>	Community guidelines and Terms of Service should clearly outline social media platforms' policy on disinformation, including definitions of what constitutes a violation and the corresponding consequences.
<b>Bot Takedowns</b>	Social media platforms should remove malicious bot accounts, botnets, or coordinated inauthentic behavior to ensure that only organic human activity is reflected in various measures of popularity, authority, and influence on social media. They are also encouraged to consider preventative measures that can include authenticity verification measures to help prevent these accounts from being created.
<b>Factchecking</b>	Social media platforms should have a factchecking process in place, preferably supported by an accredited third party that acts according to independently-established industry standards when labeling disinformation.
<b>Labelling</b>	Social media platforms should create standardized guidelines for a transparent, coherent, and multilevel labeling system dealing with (1) identified disinformation content (potentially with a ranking); (2) sponsored content (including political advertising); and (3) disinformation actors (including machine and human).
<b>Political Advertising</b>	Social media platforms should take a number of steps to clearly label sponsored content (including political advertising), including requiring verification from the sponsor and having a minimum data reporting requirement on their ad revenue streams. Along these lines, platforms should especially increase their support of current ad repositories to aid researchers. Second, platforms need to increase their oversight over political advertising, as well as limit the targeting capabilities for political advertising.
<b>Verified Information Features</b>	Social media platforms are encouraged to actively use verified information features, such as links to and pages of factchecked information that debunk disinformation during concerted campaigns or focal points, such as elections or the COVID-19 pandemic. Platforms should also be encouraged to apply such features to other societal issues, such as climate science denial, based on independent disinformation threat assessments. Finally, in exceptional circumstances, a platform should consider using its advertising algorithms to target victims of disinformation with verified information to actively debunk falsehoods.
<b>Algorithms and automated content moderation</b>	Social media platforms cannot rely on artificial intelligence alone for their moderation but must employ human moderators that are familiar with the local context and language, as well as establish efficient appeal procedures. Platforms need to be transparent about how algorithms work to both suggest and promote content, as well as how they are used in content moderation.
<b>Community Reporting and Remediation</b>	Social media platforms should have a dedicated community reporting mechanism for disinformation, take measures to timely mitigate reported disinformation and be transparent about their process. Similarly, platforms should guarantee that individuals have the ability to appeal a decision by the platform.

### 3. Develop a coregulation model to advance the standards from formulation to implementation

Based on a review of the strengths and weaknesses of the existing regulatory regimes (see Annex III), a coregulation model is proposed. It retains industry leadership in setting the standards but goes one step further than the current self-regulation model in introducing enforcement and non-compliance mechanisms, as it backs up the industry standards with a statutory layer and with independent oversight. The regulator would establish the high-level principles, while a multistakeholder body consisting of representatives from government, industry and civil society translates these high-level principles into standards codified in an industry charter. The charter does not have to start from scratch but can build on the norms, standards and KPIs provided in this report and elsewhere. The social media platforms implement these standards, and an independent oversight board monitors and reports on the performance of the platforms on the basis of which the regulator can decide to issue penalties to social media platforms.

## 4. Establish a Disinformation Information Sharing and Analysis Center (DISINFO-ISAC)

Public-private Information Sharing and Analysis Centers (ISACs), sometime known as a Warning, Advice, and Reporting Point (WARP), have formed the backbone of national cybersecurity efforts for nearly 20 years. Today, cybersecurity would be unthinkable without them. Building on this experience, a Disinformation ISAC (DISINFO-ISAC) is proposed as a distinct instrument to facilitate information exchange primarily between social media platforms, but also with government agencies, and civil society, as well as to improve the capacity of smaller industry members.

ISAC members would primarily track, label and share threat information on both disinformation content and delivery agents according to their veracity, effectively providing a 'fake ratio', and exchange their classifications according to a mutually intelligible format. That format can be based on current threat intelligence language used in cybersecurity (such as the STIX/TAXII and Snort standards) that allows threat intel to be communicated in a way that it is understandable to all humans (irrespective of language) and machine users.

Traditionally, a major part of an ISAC's taskset is supporting smaller members as well as educating external stakeholders. It provides them with crucial threat intelligence and functions as a discussion forum to use as they see fit. As a result, those smaller members that lack resources would be able to take less drastic moderation solutions when responding to disinformation. Civil society organizations and academia could be encouraged to give input into the DISINFO-ISAC, and provide an ability to audit and, if necessary, appeal false classifications. Such a remediation process for falsely tagged content can be done by the ISAC, which passes on complaints to the responsible body and follows up if there is a lack of response, and would therefore have a limited 'ombudsman' type function.

The DISINFO-ISAC would not replace current ad-hoc threat information exchanges between platforms. While such informal exchanges can be very effective, collaboration also needs to be seen by outsiders, rather than relying on the judgement of individual companies. It would be an industry-led initiative but through its involvement of civil society and government stakeholders, the ISAC does not only show it can be a forum for effective collaboration, but also ensures that such collaboration is *seen* by outsiders. It would thereby improve public-private cooperation and trust, as well as inform related government-to-government initiatives. By involving the EEAS and the Canadian Secretariat of the G7 Rapid Response Mechanism, it can inform the EU Rapid Alert System (RAS) and G7 Mechanism, allowing them to receive real-time threat updates from the social media platforms. Government partners can then combine that threat information with their own classified information and operationalize this into mitigation guidance or otherwise useful information that non-cleared partners can use. Taken together, the DISINFO-ISAC forms a crucial step towards a common crisis-management mechanism or "break-the-glass protocol" in the fight against disinformation.

These recommendations come at a time when the European self-regulatory approach towards social media companies' responsibility is shifting towards coregulation. The proposed counter-disinformation standards and key performance indicators, as well as the recommendations for accountability and multistakeholder engagement in coregulation are

therefore timely. Most notably the DISINFO-ISAC would be a major contribution to European norm-setting and illustrates that a coregulation model could be made very tangible. Social media platforms and EU institutions are therefore called upon to consider these and the proposals made elsewhere to strengthen the European Democracy Action Plan, improve the responsibility of social media platforms in countering disinformation, and strengthen their cooperation with other stakeholders.

# 1 Introduction

Disinformation continues to dominate the headlines as one of the major political challenges that liberal democracies face today. In recent years, disinformation campaigns have exacerbated existing political polarizations with effects as diverse as inciting mass violence against the very institutions of democracy itself, as well as discrediting measures against the COVID-19 pandemic. They have undermined trust in democratic processes and even had a significant effect on elections and plebiscites across Europe and North America.

Governments struggle to find effective ways to counter this threat. This is in no small part due to the enormously complex nature of the issue of disinformation, the wide range of actors involved, and the numerous dilemmas it presents within and across themes such as national and international security, democracy and human rights, cybersecurity, and Internet governance. Social media platforms also struggle with their position at the frontline of the ‘infodemic’. They see themselves increasingly forced to take action to keep their platforms from being used to spread disinformation. Most notably, the EU has done so through the European Democracy Action Plan, the Code of Practice on Disinformation, and is addressing platform responsibility more broadly through the Digital Services Act (DSA) and Digital Market Act (DMA). Disinformation is just the latest addition to a wide range of other issues – including terrorist content and online child abuse – on the basis of which these companies have been nudged to increasingly moderate content. Disinformation, however, is not explicitly illegal under international law. Rather than being a widely-agreed legal term, it is a policy term that has major freedom of speech implications, and appears to be a much more contentious issue.

Increasingly, governments are urging platforms to take more responsibility ranging from self-regulation initiatives to binding laws. At the same time, some politicians and civil rights organizations criticize these platforms for censorship and have proposed laws that aim to prevent platforms from removing too much content. In some cases, governments have adopted legal measures, sometimes in conjunction with hate crime legislation, to compel these platforms to self-moderate. Such legislation – at both the national and international levels – can be abused to impede fundamental human rights. Authoritarian regimes have used such legislation to crack down on free speech and suppress any dissenting or critical voices against their regime.

As of yet, there is no ‘gold standard’ in combating disinformation. Liberal democracies seem to vacillate between *over* and *under* regulation, neither of which is without risks. As with other emerging security challenges, such as cybersecurity and hybrid conflict, disinformation is a field that lacks consensus for a common unilateral legal solution due to its complex and layered nature. Nonetheless, ‘rules of the road’ to guide responsible behavior are needed. In the cyber realm, rather than crafting new treaties, the international community has thus far preferred to establish norms – voluntary, legally non-binding commitments that reflect a common standard of acceptable and proscribed behavior. These norms accompany and expand on existing legal understandings rather than attempt to craft new law.

In the 2020 report *From Blurred Lines to Red Lines – How Countermeasures and Norms shape Hybrid Conflict*, we offered recommendations on how norms and countermeasures can shape

As of yet, there is no ‘gold standard’ in combating disinformation. Liberal democracies seem to vacillate between over and under regulation, neither of which is without risks.

---

the behavior of hybrid threat actors and raise costs for an attack.<sup>1</sup> This report builds on those recommendations and asks what kind of norms can be developed, by whom and for whom to counter disinformation? And finally, how can these norms be advanced?

This report addresses these questions by means of a government-to-government norm as well as through industry norms. Both avenues take place at the international “Whole of System” level. This goes beyond the “Whole of Government” or “Whole of Nation” approach and includes a wider community of interest from government, industry and civil society (see Textbox 1). This approach is applied at two levels. First, global government-to-government norm development in which states, as the leading stakeholder group and primary audience, are urged to show restraint when it comes to disinformation operations. However, this comes with great risks to civil rights and potentially endangers the existing multistakeholder approach of Internet governance. Furthermore, governments only make up one of three actor groups of the wider information environment. Industry, in particular social media platforms, is arguably the most prevalent actor group in this space, and has developed its own norms, standards, and best practices against disinformation. The second approach therefore looks at norm development focused on social media platforms, culminating into an industry charter of standards against disinformation. It does so at the European level to follow up and inform EU efforts, including through the European Democracy Action Plan and the upcoming Digital Services Act. Finally, a coregulation model and a disinformation-focused Information Sharing and Analysis Centre (DISINFO-ISAC) are proposed, in which the private sector, EU institutions, and civil society operate side-by-side in advancing the standards from formulation to implementation.

#### **Textbox 1: Whole of Government, Whole of Nation and Whole of System approach.**

A national counter-disinformation strategy needs to have a national Whole of Government and Whole of Nation approach, as well as an international Whole of System approach.<sup>2</sup> The esoteric nature of the many individual mandates involved in countering disinformation naturally leads to ‘stovepiping’ in narrowly defined government organizations. The reality of these different mandates is that they are each dealt with by different organizational groups within government, but also within the non-state sector both nationally and internationally.

**The Whole of Government approach** depends on successful *coordination* between government agencies at the central, state and/or local level.

**The Whole of Nation approach** depends on successful *cooperation* between national state and non-state actors, their national civil society and industry partners. This approach is particularly relevant for issues like counter-disinformation and cybersecurity in which the government is not the predominant stakeholder group.<sup>4</sup>

**The Whole of System approach** depends on *collaboration* with a wide range of state and non-state partners at the international or regional level, whether it is through binding treaties, norms, or non-governmental agreements between non-state actors. EU-led approaches have sometimes been entitled “Whole-of-the-Union” while in NATO the formulation “Whole-of-Alliance” is used.

All three approaches are required. They just need to inform each other, making sure there is sufficient mutual awareness and a clear link between the national and international strategies.

1 Louk Faesen, Tim Sweijjs, Alexander Klimburg, Conor MacNamara and Michael Mazarr, *From Blurred Lines to Red Lines: How Countermeasures and Norms Shape Hybrid Conflict*, (The Hague: The Hague Centre for Strategic Studies, September 2020).

2 Inspired by the Cybersecurity context; see: Alexander Klimburg (Ed.), *National Cyber Security Framework Manual*, (Tallinn: NATO CCD COE Publications, 2012).

3 The Dutch Cabinet position, for example, states that responding to disinformation is not primarily a task of the government, unless it involves illegally acquired, unlawful or disruptive information, or in case of a threat against the political and economic stability or national security. This implies that other – non-state – actors, like platforms and media, play a crucial role in countering disinformation.

The report proceeds as follows:

- **Chapter 2** introduces the issue of disinformation, the definitions used, and shows the complexity of the counter-disinformation regime complex and its many actors involved.
- **Chapter 3** defines norms and explains the difference between the norms established by governments – big N Norms – and the best practices or standards set by non-state actors – small n norms.
- **Chapter 4** explores the viability, benefits and risks of a big N Norm for government restraint in disinformation, as well as possible avenues for its propagation.
- **Chapter 5** explains why companies at the top of the Tech Stack are best placed for moderating online content and conduct, and identifies eight small n norms or standards against disinformation as part of an industry charter for social media platforms.
- **Chapter 6** proposes a European coregulation model and a DISINFO-ISAC to advance the industry charter of standards and facilitate public-private cooperation across the counter-disinformation regime complex.
- **Chapter 7** concludes and offers policy recommendations for the advancement of the industry charter, the coregulation model, and the ISAC.
- **Annex I** offers a detailed overview of the current best practices of the major social media companies in countering disinformation.
- **Annex II** includes two case studies: the Global Internet Forum to Counter Terrorism offers lessons learned from centralized industry cooperation on content moderation, and the EU Code of Practice on Disinformation on the self-regulatory approach to counter disinformation.
- **Annex III** provides an overview of the various regulatory regimes, from self-regulation, to coregulation and statutory regulation.
- **Annex IV** describes the important role of ISACs in cybersecurity information exchange, and explains its overall functions, actors and overall roles, and possible organizational structures.
- **Annex V** includes the list of interviewees that were consulted for feedback on the proposals in this report.

## 2 Introduction to disinformation: definitions and the regime complex

Today, many authoritarian governments see the Internet not only as an opportunity but also as a potential threat to their regime.

From the Internet's very beginning, the online information environment has been loosely governed.<sup>4</sup> While this partially has been a reflection of early interpretations of US law (in particular the famous Section 230 of the US Communication Decency Act), it is also a reflection of how rapidly technological developments have outpaced legislation. While the loose governance structure encouraged an unprecedented exchange of ideas, communication, economic and social benefits, it has also created risks and challenges. Today, many authoritarian governments see the Internet not only as an opportunity but also as a potential threat to their regime. After all, the Internet has become an area in which countries can advance their foreign policy goals. This has created a growing sense of concern in the international community and the public at large, especially as authoritarian regimes have attempted to use international forums to propose norms and laws justifying more government control over the Internet (further discussed in Chapter 4). As such, there is growing and widespread demand for liberal democracies to develop and foster better and more explicit definitions and governance structures involved in countering disinformation.

This chapter starts by offering definitions of disinformation, and its relation to other forms of information disorder and influencing operations, followed by a non-exhaustive mapping of the existing 'counter-disinformation regime complex'. The latter visualizes the complex nature of the ecosystem and its myriad actors and initiatives involved in countering disinformation. This constitutes a first step towards identifying where actors with similar interests and positions can mutually reinforce each other and where duplication of efforts can be avoided.

### 2.1 Definitions

Many of the stakeholders involved in mitigating disinformation use their own terminologies to encapsulate the problem, including information warfare, influence operations, hybrid

<sup>4</sup> For a definition of the information environment, see US JP-3-12 Cyberspace Operations: "The information environment is the aggregate of individuals, organizations, and systems that collect, process, disseminate, or act on information.", Joint Staff. "Joint Publication 3-12: Cyberspace Operations." JCS.mil, (8 June, 2018): [https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3\\_12.pdf](https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_12.pdf).; Cyberspace is considered to be part of the information environment, and is defined by the Netherlands Military Cyberspace Doctrine in the same way as the NATO AJP 3.20 allied Joint Doctrine for Cyberspace Operations: "The global domain consisting of all interconnected communication, information technology and other electronic systems, networks and their data, including those which are separated or independent, which process, store or transmit data." Ministry of Defense of the Netherlands, "The Netherlands Armed Forces Doctrine for Military Cyberspace Operations". Dutch Defense Cyber Command, (June 2019).



warfare, coordinated inauthentic behavior, or computational propaganda. One of the most widely-used definitions of disinformation explains it as “information that is false and deliberately created to harm a person, social group, organization or country.”<sup>5</sup> From a governmental perspective, disinformation is not a legal but a policy term. After all, disinformation is not illegal under current international law and most European have not regulated it, with the notable exception of the French Fake News Law.<sup>6</sup> The European Democracy Action Plan, of which the EU Code of Practice on Disinformation is a part, uses a similar definition but includes the “intention to deceive or secure economic or political gain.”<sup>7</sup> These definitions rest on two core components: the *falsehood* and the *organized intent to deceive or harm*. If one was to plot this on a scale (see Figure 1), misinformation or “information that is false, but not created or distributed with the intention of causing harm” would be placed closer to falseness, while malinformation, or “information based on reality, used to inflict harm on a person, organization or country” would be placed closer to the intent to harm, much like influence operations and foreign interference.<sup>8</sup>

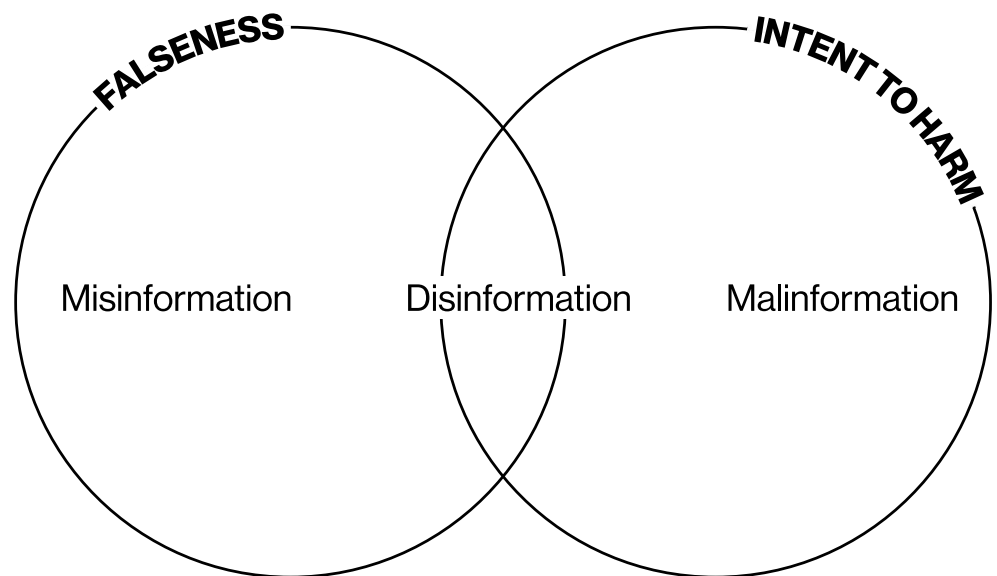


Figure 1: The Relationship between Disinformation, Misinformation, and Malinformation.<sup>9</sup>

5 Claire Wardle and Hossein Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policymaking*, (Strasbourg: Council of Europe, 2017), 30. This definition formed the basis for the definitions used by the Council of Europe and the 2018 report from the independent European High-level Group on fake news and online disinformation, which has then, in turn, been used as a definitional for the EU Code of Practice on Disinformation. This definition also aligns with definitions used by other governments. For instance, as reported in the 2019 Disinformation and ‘fake news’ report, the UK government uses a definition that closely mirrors this definition.

6 “Against information manipulation”, France, November 20, 2018, <https://www.gouvernement.fr/en/against-information-manipulation>

7 European Commission, *On the European democracy action plan* (Brussels: 2020), 18. This definition largely corresponds to the official definition of the Netherlands, which described it as: disseminate misleading information, mostly covertly, with the intent to damage the public debate, democratic processes, the open economy and national security.

8 Wardle and Derakhshan, *Information Disorder*, 18.

9 Claire Wardle, “Understanding Information disorder,” First Draft, September 22, 2020, <https://firstdraftnews.org/long-form-article/understanding-information-disorder/>.

One of the most widely-used definitions of disinformation explains it as “information that is false and deliberately created to harm a person, social group, organization or country”



Overall, tackling conduct issues is less controversial because it avoids an arbitration over what is “true” and “false” information and presents fewer risks to online free speech.

Counter-disinformation efforts generally focus on either *content* or *conduct*. The latter focuses on malign behavior used to disseminate false content rather than the content itself and includes issues such as countering coordinated inauthentic behavior, impersonation, hack and dump operations, the use of bots and spam. Overall, tackling *conduct* issues is less controversial because it avoids an arbitration over what is “true” and “false” information and presents fewer risks to online free speech. The focus on *conduct* also broadens the scope of disinformation, as reflected in the European Commission’s *Guidance on Strengthening the Code of Practice on Disinformation*. This report takes a similar approach and primarily looks at disinformation in the narrow sense as previously defined, while simultaneously adopting a broader interpretation. This broad interpretation takes into consideration the close link of disinformation to other forms of information disorder, such as misinformation, influence operations and foreign interference.<sup>10</sup> Acknowledging the nuanced differences between these terms, foreign interference and covert influence operations are of particular relevance for Chapter 4 which proposes an intergovernmental norm to counter disinformation on the basis of noninterference, while the growing threat of misinformation could also be countered by some of the proposals made in Chapter 5 for social media platforms to counter disinformation.

## 2.2 The Regime Complex for Countering Disinformation

A number of actors and initiatives, ranging from regulators to social media platforms to civil rights organizations, act together in so-called “regimes” to determine the governance course of various complex multistakeholder topics. Such topics include for example environmental policy making or the management of Internet resources, commonly referred to as Internet governance. Regimes can be defined as implicit and explicit “principles, norms, rules and decision-making procedures around which the actor expectations converge in a given issue-area.”<sup>11</sup> Each regime effectively engages in governance within a specific thematic area, defining accepted standards, policies, or laws. The process of governance is encapsulated in a number of different assumed responsibilities and activities – ranging from non-state-led processes (e.g. technical and industry initiatives) to state-led processes on international security issues (e.g. within the United Nations and regional organizations). This is no different in the fight against disinformation.

Together, this “galaxy” of initiatives forms the “Countering Disinformation Regime Complex” (see Figure 2). One of its important features is that it is very much multistakeholder in its composition – the government, the private sector and civil society (which includes the technical community as well as academia and NGOs) all play a role – very often together. Given that many of the regimes are unique and autonomous, but also often work at odds with each

<sup>10</sup> The European Commission tends to use the definitions for Foreign Interference and Information Influence Operations as identified in their 2020 document, *On the European Democracy Action Plan*. There, they write that Information Influence Operations are “coordinated efforts by either domestic or foreign actors to influence a target audience using a range of deceptive means, including suppressing independent information sources in combination with disinformation,” and Foreign Interference in the Information Space are “coercive and deceptive efforts to disrupt the free formation and expression of individuals’ political will by a foreign state actor or its agents.” These definitions themselves are heavily influenced by the work of James Pamment. European Commission, *On the European democracy action plan*, 18; James Pamment, *The EU’s Role in Fighting Disinformation: Taking Back the Initiative* (Washington DC: the Carnegie Endowment for International Peace, 2020), [https://carnegieendowment.org/files/Pamment\\_-\\_Future\\_Threats.pdf](https://carnegieendowment.org/files/Pamment_-_Future_Threats.pdf).

<sup>11</sup> Stephen D. Krasner, “Structural Causes and Regime Consequences: Regimes as Intervening Variables,” *International Organization* 36, no. 2 (Spring 1982). <https://www.jstor.org/stable/2706520>.

other, it is becoming evident that finding coherence amongst them is a critical step in being able to define actionable measures, irrespective of whether they are mainly technical, legal, or political. The regime complex visualization does not attempt to capture one dominant national approach described in the introduction. Instead, it illustrates the existence of a variety of approaches.

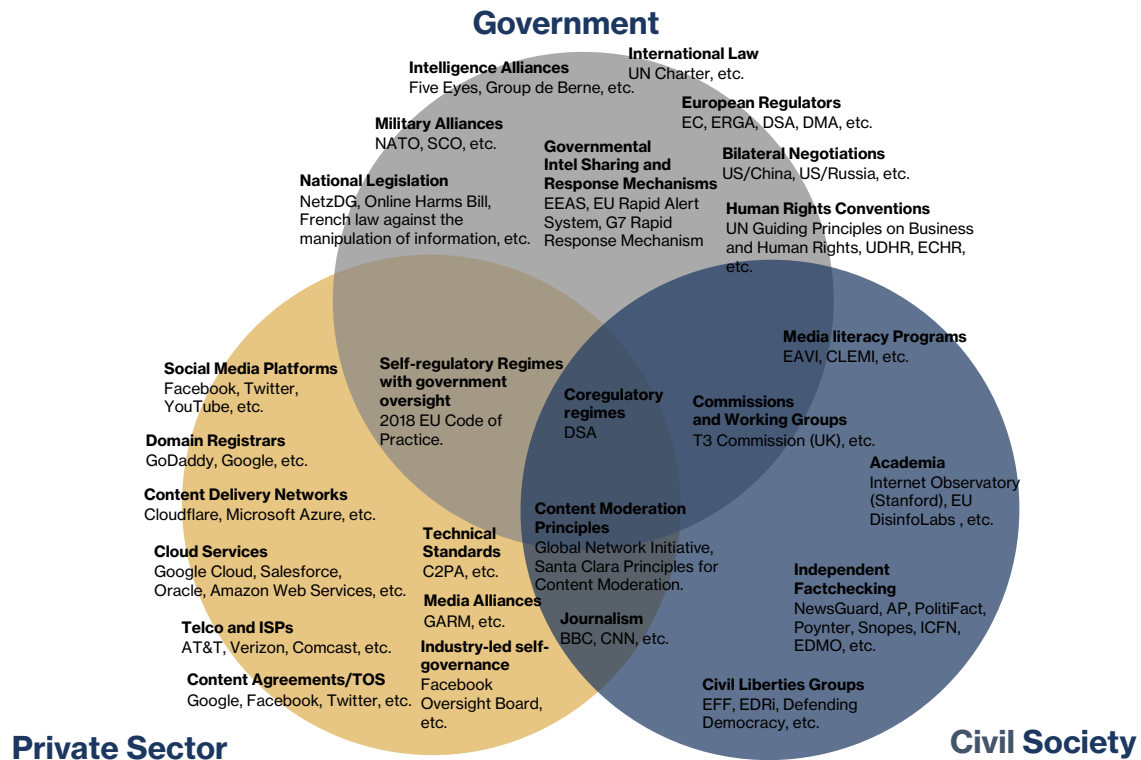


Figure 2: The Countering Disinformation Regime Complex.<sup>12</sup>

Despite being a field still in its infancy, the range of relevant stakeholders and actors in the information environment dealing with disinformation is large and complex. As demonstrated in Figure 2, stakeholders can be plotted on a Venn diagram with the three main types of stakeholders being government, private sector, and civil society.<sup>13</sup> Government initiatives include national laws, actions, and task forces which combat disinformation, as well as international and inter-governmental initiatives. Private sector stakeholders are wide and diverse, spanning from social media platforms to the less well-known parts of the Internet, such as Internet Service Providers (ISPs) and Content Delivery Networks (CDNs). Finally, civil society stakeholders are vast and varied, typically focusing on bringing attention to gaps in the efforts of the private sector of the government. Compared to the cyberspace regime complex, only a small number of initiatives or actors overlap with multiple stakeholder groups. Most notably are multistakeholder consultations and negotiations held by international bodies, and the emergence of regulatory regimes, such as the self-regulatory 2018 EU Code of Practice on Disinformation.

<sup>12</sup> Derived from Alexander Klimburg and Louk Faesen, "A Balance of Power in Cyberspace," in *Governing Cyberspace: Behavior, Power, and Diplomacy*, ed. Dennis Broeders and Bibi van den Berg (London: Rowman & Littlefield, 2020).

<sup>13</sup> The underlying idea behind the mapping of the regime complex is not to provide an exhaustive list of actors or a definitive analysis, but they visualize the complex nature of the disinformation ecosystem, its actors, and initiatives, each with their own standards and norms, that can partly overlap.

## 2.3 Key Takeaways

There are nearly as many definitions for disinformation as there are stakeholders. Disinformation can be defined as “information that is false and deliberately created to harm a person, social group, organization or country,”<sup>14</sup> and functions as a basis for this report. It rests on the notion of falsehood and the intent to harm, and primarily focuses on concerted interference efforts from foreign state actors.

The ability of governments to successfully manage the threat is hampered by the dominant role of non-state actors as attacker, victim and carrier of disinformation. Governments only constitute one of three stakeholder groups in the wider counter-disinformation ecosystem. Given this complex landscape, it is unlikely there can be a unilateral legal or normative solution that works across the entire counter-disinformation regime complex, from government to civil society and industry.

---

14 Ibid 5.

# 3 Norms

## 3.1 Introduction by Dennis Broeders

Disinformation combines the domestic and international: foreign information operations are only effective if they land on fertile ground in the target state

There is no easy way for liberal democracies to address the problem of disinformation. Domestically, the problem and the possible solutions intersect with core beliefs about freedom of speech and the openness of the (digital) public sphere. Internationally, they rub uncomfortably against notions of 'information security' and 'cyber sovereignty' that are used by adversarial states like Russia and China to justify surveillance and control of their own digital domestic public sphere. Disinformation combines the domestic and international: foreign information operations are only effective if they land on fertile ground in the target state. They play on societal divisions and nativist tendencies making it hard to separate out the 'foreignness' of disinformation,<sup>15</sup> and force western societies to look in the mirror. The problem is, however, growing – election interference and Covid-19 related disinformation for example – and liberal democracies are increasingly under pressure. This report identifies two routes to address the problem: a diplomatic route through Big N norms and a small n route going through technical and corporate gatekeepers. Both routes are beset by dragons.

The playing field for Big N norms is the UN First Committee where states have been discussing 'responsible state behaviour in cyberspace' in consecutive United Nations Groups of Governmental Experts (UN GGE) and in the more recent Open Ended Working Group (OEWG) processes. These processes originated with a 1998 UN resolution tabled by Russia out of fear for foreign interference in the Russian digital public domain.<sup>16</sup> Ironically, what Russia then feared most, it now does best. As defenders of the idea of an open and free Internet and fundamental rights such as freedom of speech, the like-minded countries rejected the Russian premise and terminology of 'information security' and focused the GGE on cybersecurity rather than information and content. Moreover, they side stepped the idea of a new binding treaty and instead pushed a new interpretation of soft law, so-called "norms". As successful as the like-minded group of democracies was in turning the GGE away from content as a security issue and new international law and treaties, problems with content seemed to grow in their own countries as well. The challenge of disinformation and foreign influence operations however steers them back towards the language of 'information security' that has now been cornered by authoritarian states.

Another problem with Big N norms is that a norm against disinformation would have to be a negative norm: thou shalt not conduct foreign influence operations. Such a norm would require states to refrain from doing something and would either have to be internalized by all relevant members of the community (self-restraint) or vigorously policed by other members of the community (external restraint). A negative norm itself cannot be 'implemented', at best the international community can implement cooperative measures, CBMs and support

15 Hedvig Ördén and James Pamment, *What Is So Foreign About Foreign Influence Operations?*, (Washington DC: Carnegie Endowment for International Peace, 2021). [https://carnegieendowment.org/files/Orden\\_Pamment\\_ForeignInfluenceOps2.pdf](https://carnegieendowment.org/files/Orden_Pamment_ForeignInfluenceOps2.pdf)

16 UN General Assembly, "United Nations General Assembly Resolution 53/70", *United Nations*, January 4, 1999. <https://undocs.org/A/RES/53/70>

capacity building to support a norm. The most sustainable anchoring of any norm is through internalization, but the deck is stacked against that happening. Disinformation campaigns have turned out to be an effective and low-cost way for Russia, and increasingly other authoritarian regimes, to achieve foreign policy goals. Stoking up and reinforcing discord in Western liberal democracies threatens to discredit democracy as an alternative model to authoritarian rule, muddles the domestic debate and leads attention away from other international disputes and tensions. Moreover, liberal democracies cannot respond in kind for operational and political reasons. Operationally, it's not a level playing field. Authoritarian regimes have much more control over their national information spheres, giving them a defensive advantage, while liberal democracies' open information spheres and spaces for public debate provide vulnerable targets. More importantly, conducting covert information operations makes liberal democracies vulnerable to charges of hypocrisy. It's hard to defend the integrity of information by fighting the integrity of information. Fighting fire with fire is likely to get democracies burned.<sup>17</sup>

The high road would be to address the issue under international law. To some extent the problem is elusive because states lack the legal language to define it as a problem, or lack the political will to apply legal language to define it as a problem. This report rightly points towards sovereignty and non-intervention as possible avenues to flag disinformation as a violation of international law. Legal scholars such as Schmitt,<sup>18</sup> Tsagourias,<sup>19</sup> and Ohlin,<sup>20</sup> have done great work on trying to connect the problem of disinformation with the legal principles of non-intervention and national self-determination, flagging its 'covertness' as being legally significant. However, given the legal requirements for applying these international law principles, such as proving 'coercion', disinformation often gets narrowed down to 'election interference'. In the context of the UN's Big N norms, it gets narrowed down even further. The general mood towards UN norms in the last years have been to 'add a layer of understanding' to the 11 norms of the 2015 GGE report, rather than draft new norms. This makes the 2015 norms on the protection of critical infrastructure the most readily available framing, narrowing election interference down further to interference with election *infrastructure*, effectively cutting out information operations and disinformation.<sup>21</sup> The 2021 UN GGE consensus report flags "States' malicious use of ICT-enabled covert information campaigns to influence the processes, systems and overall stability of another State" as an emerging problem in the threat section, but does not revisit it in the norms or international law sections.<sup>22</sup>

It's hard to defend the integrity of information by fighting the integrity of information. Fighting fire with fire is likely to get democracies burned.

17 Dennis Broeders, "Creating Consequences for Election Interference," *Directions. Cyber Digital Europe*, May 15, 2020. <https://directionsblog.eu/creating-consequences-for-election-interference>

18 Michael Schmitt, "'Virtual' Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law," *Chicago Journal of International Law* 19, no. 1 (2018):30–67. <https://chicagounbound.uchicago.edu/cjil/vol19/iss1/2>

19 Nicholas Tsagourias, "Electoral Cyber Interference, Self-Determination, and the Principle of Nonintervention in Cyberspace." In *Governing Cyberspace: Behaviour, Power and Diplomacy*, edited by Dennis Broeders, and Bibi van den Berg. (London: Rowman & Littlefield, 2020) 45–63.

20 Jens David Ohlin, "Election Interference: The Real Harm and The Only Solution", *Cornell Legal Studies Research Paper* No. 18-50 (2018). <https://ssrn.com/abstract=3276940>; Jens David Ohlin *Election Interference. International Law and the Future of Democracy*. (Cambridge: Cambridge University Press, 2020). doi:10.1017/9781108859561

21 Dennis Broeders, "The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: a mid-process assessment", *Journal of Cyber Policy* (2021). DOI: 10.1080/23738871.2021.1916976

22 United Nations Group of Governmental Experts, "Report of the Group of Governmental Experts on Advancing responsible State behaviour in cyberspace in the context of international security", *United Nations*, May 28, 2021 (advance copy). <https://front.un-arm.org/wp-content/uploads/2021/06/final-report-2019-2021-gge-1-advance-copy.pdf>

Co-regulation can help to avoid the problem of creating 'norms formulated by one set of actors but expected to be executed by another', while adding a serious mechanism of oversight and accountability can help to go beyond the often empty self-regulation that we have seen before in this domain

Because of the trouble with Big N norms, this report turns to small n norms as an alternative route to deal with disinformation. Small n norms have some distinct advantages: they deal with the trouble (disinformation) but do not necessarily point the finger at the troublemaker (an adversarial state). From an international perspective it is a fire brigade logic (put out the flames, and improve fire safety) rather than a law enforcement logic (find and arrest the arsonist). Given the deep divisions in the United Nations about attribution this may indeed be a more viable option.<sup>23</sup> Another advantage is that solutions can be scaled down to the region. Instead of chasing after consensus among all UN member states, the challenge in this report becomes creating a viable multistakeholder governance regime at the level of the European Union. However, this route is also not unproblematic because at the core of any counter disinformation policy is content control, although some forms of 'conduct control' advocated in this report – such as botnet takedowns – at least partially steer around that problem. Firstly, in liberal democracies the unease about disinformation is paralleled by unease about (social media) platforms shaping the digital public sphere, and turning to them for content control would only increase their influence. Secondly, as Madeline Carr argued about public-private cooperation in cybersecurity: public interests and private interests seldom align in full as corporations do not think in terms of common public goods.<sup>24</sup> While social media platforms may be concerned about disinformation, it spreads virally on the back of the algorithms that govern the attention economy on their platforms. On platforms, viral spread is not a bug, it is a feature. Thirdly, big tech platforms do not usually align with foreign policy goals either. Market capture often necessitates a "when in Rome" mentality to the requests and requirements for content moderation in autocratic states that liberal democracies think unsavory. This means that multistakeholder solutions should be based on realistic assessments of the interests and behavior of the stakeholders involved. This report offers a way forward. Co-regulation can help to avoid the problem of creating 'norms formulated by one set of actors but expected to be executed by another'<sup>25</sup>, while adding a serious mechanism of oversight and accountability can help to go beyond the often empty self-regulation that we have seen before in this domain. Verification may help to build trust among stakeholders. However, small n, regional solutions such as those proposed in this report still have to tread carefully in light of fundamental rights and the international debate about content and information, not in the least to counterbalance the instrumental use of the terminology by authoritarian states.<sup>26</sup>

## 3.2 Norms Primer

Given the complexity of the disinformation ecosystem, consensus on a common unilateral legal solution that all actors would agree to is highly unlikely. As such, tools like norms can be especially useful. Their dynamic nature constitutes a more flexible and agile regulatory alternative to formalized binding laws to address emerging security threats such as disinformation

23 "A Guide to the UN GGE", narrated by James Lewis and Chris Painter, Inside Cyber Diplomacy, *Center for Strategic & International Studies*, June 11, 2021. , <https://www.csis.org/node/61229>

24 Madeline Carr, "Public-private partnerships in national cyber-security strategies", *International Affairs* 92, no. 1 (January 2016), 43-62. <https://doi.org/10.1111/1468-2346.12504>

25 For further reference, please see the section of this report *Conclusion and Recommendations*.

26 The fight against disinformation through platforms is ultimately about content control and will be framed by Russia and China as the western version of information security. They will use it as a justification for their domestic control and surveillance mechanisms and as a shield against Western criticism. Moreover, the issue of controlling Western platforms because of the 'damaging content', i.e. critical content, they spread has already surfaced in the UN in the contributions of states like Venezuela, Iran and Zimbabwe to the OEWG (see Dennis Broeders, Fabio Cristiano, and Daan Weggemans, "Too Close for Comfort: Cyber Terrorism and Information Security across National Policies and International Diplomacy", *Studies in Conflict & Terrorism*, (2021), p.15. DOI: 10.1080/1057610X.2021.1928887)



A norm is generally defined as “a collective expectation for the proper behavior of actors with a given identity”

by establishing internationally agreed-upon thresholds for behavior. They are also not exclusively set by governments but can also be established by non-state actors from industry or civil society. Their voluntary nature also allows wide agreement on a broader set of rules of the road without getting bogged down in thorny legal negotiations that typically take many years to complete.

A norm is generally defined as “a collective expectation for the proper behavior of actors with a given identity”, consisting of four core ingredients: (i) identity (the who); (ii) propriety (the how); (iii) behavior (the what and where); and (iv) collective expectations (the why).<sup>27</sup> There may be differences in the usage of the term “norms” depending on the ingredients, such as regulative, constitutive and proscriptive norms.

Norms are not static tools of diplomacy, nor do most norms emerge spontaneously. They are the result of dedicated work by actors to promote a new standard of behavior for reasons ranging from self-interest and values to ideational commitment. These actors are the norm entrepreneurs that may be any group of actors – states, international or regional organizations, companies, transnational NGOs, or other interest-groups. Successful norms are often supported by a broad coalition of actors including not only states but also industry civil society organizations. Developed by Martha Finnemore, the norm lifecycle catalogs the development and propagation of norms across three stages: norm emergence, norm cascade and norm internalization (see Table 2).<sup>28</sup>

Table 2: The Three Stages of the Norm Lifecycle: Norm emergence, norm cascade, norm internalization

Stage 1: Norm Emergence	Stage 2: Norm Cascade	Stage 3: Norm Internalization
Habit and repetition alone – particularly when they go unchallenged – create norms. Alternatively, it can be a dedicated effort by a norm entrepreneur, who has the first-mover advantage of <i>framing</i> a norm within a preferential context and <i>linking</i> it to other pre-existing norms, which not only increases its credibility and urgency but also anchors the norm within the values and interests of the entrepreneur.	Once a sufficient number of actors have been persuaded by the entrepreneur or even coerced into acceptance, it can trigger socialization effects, like bandwagoning or mimicry, on the remaining hold-outs, accelerating the norm towards widespread acceptance. This process is accelerated when the norm is grafted to organizational platforms.	When a norm is internalized it is ‘taken for granted’ and no longer considered ‘good behavior’; rather it becomes a foundational expectation of acceptable behavior by the international community. Once internalized, a norm shapes the interests of states rather than vice versa. Internalized norms however continue to evolve as the interests, context, identity, and propriety change around them.

Norm entrepreneurs typically use two key processes to establish their proposed norms within the first stage of norm emergence: *framing* and *linking*. For *framing*, norm entrepreneurs *frame* the norm within a specific context using language and interpretations that are intrinsic to their perceptions, interests and values.<sup>49</sup> Norms are often framed in several manners to appeal to several different audience’s interests. Next, proposed norms are often *linked* to existing norms or impactful issues that can attract attention and resources or reinforce it. Ultimately, the strategic selection of the context through framing and linking will determine the reach and pathway of the norm, its strategy, target audience, and the tools of influence. Three tools of influence are identified that contribute to norm cascade (wide acceptance) and internalization (deep acceptance) (see Table 3).<sup>29</sup>

27 Peter J. Katzenstein, *The Culture of National Security: Norms and Identity in World Politics* (New York: Columbia University Press, 1996).

28 Martha Finnemore and Kathryn Sikkink, “International Norm Dynamics and Political Change,” *International Organizations* 52, no. 4 (1998): 887-917. <https://www.jstor.org/stable/2601361?seq=1>.

29 The tools of influence for norm promotion can also be used for norms to counter hybrid operations. See: Faesen et al., *From Blurred Lines to Red Lines: How Countermeasures and Norms Shape Hybrid Conflict*.

Table 3: Three tools for norm promotion: socialization, persuasion and coercion

**Socialization** leverages the shared relations and identities between actors and institutions in order to push a norm towards conformity. It includes forms of conformity based on national interests, such as rationally expressive action, social camouflage, bandwagoning, mimicry, insincere commitments to avoid stigmatization, or improved relations.

**Persuasion** can occur through cognitive means (through *linking* or *framing*) or material incentives. Persuading actors through *incentives*, such as trade agreements, is mostly a tool available to strong states as they require a vast amount of resources over a longer period of time.

**Coercion** refers to the use of negative inducements, such as sanctions, threats, and indictments to promote the norms of the strong. It largely remains a tool for strong states that have attribution capabilities and political will. When entrepreneurs face opposition from other actors in the contentious stages of the norms lifecycle, incentives and coercion can play a major role.

A norm entrepreneur should take advantage of the wider spectrum of tools and realize where they enforce their strategy or potentially crowd out other tools. Each tool comes with its own set of costs and benefits that require the entrepreneur to continuously (re)evaluate their choices based on their interests and changing contexts. Ultimately, the success of a norm rests not just in its content, but in its process: who pushes it, accepts it, and where, when, and how they do so.

For the purpose of this report, a distinction is made between “big N norms” – the regulative norms established by governments – and “small n norms” – the best practices or standards set by non-state actors. Before taking a closer look at the specific governmental and non-governmental normative initiatives on disinformation, this chapter introduces norm development in both contexts and explains how they differ from each other and what kind of challenges they pose for creating norm coherency across the counter-disinformation regime complex.

### 3.1 Big N Norms

Establishing finely delineated legal responsibilities for an immensely complicated and layered regime complex is often not possible. Since there currently is no authoritative governmental process or consensus on norms against disinformation, the process and dynamics of “big N Norms” are best explained by looking at the intergovernmental norm-setting process for cybersecurity.

In 1998, Russia introduced a resolution on information and telecommunications technology in the context of international security to the United Nations General Assembly.<sup>30</sup> This represented the first, and far from the last, time that the topic of cybersecurity was to be addressed under the auspices of the United Nations. Since then, progress has been slow. As time has demonstrated, binding legal agreements, like the one proposed by Russia, have proven to be too difficult given the definitional and ideological differences between East and West. Thus, in the absence of a clear and viable path towards a treaty-based solution, the international community instead focused on a path of norm development through the UN Group of Governmental Experts on Developments in the field of information and telecommunications in the context of international security (UN GGE).

30 UN General Assembly, “United Nations General Assembly Resolution 53/70,” (United Nations: January 4, 1999).

A distinction is made between “big N norms” – the regulative norms established by governments – and “small n norms” – the best practices or standards set by non-state actors



In 2013, the UN GGE established a consensus that International Law applies to cyberspace, just as it does to other domains.<sup>31</sup> Exactly *how* it should apply remains a matter of contention. While countries like Russia and China have long advocated a novel treaty-based approach, liberal democracies have insisted that *existing* international law needs to be the point of departure.<sup>32</sup> Since 2013, the compromise achieved by both parties has been agreed-upon norms of behavior via the consensus reports of the UN GGE adopted by the UN General Assembly, and more recently a parallel process through the UN Open-Ended Working Group on ICTs (OEWG).<sup>33</sup> Western like-minded countries opted for this legally non-binding solution as an alternative to new treaties because they maintained the liberal-democratic status quo of *existing* international law. Russia, China and their allies on the other hand, perceive these norms not as an enhanced interpretation of existing international law *per se*, but as evidence that existing law falls short and that norms are the first step towards *new* international law.

These diplomatic processes have seen some successes in establishing “Big N norms” in cyberspace. Notably, in 2015, the UN GGE presented eleven norms, which included commitments such as requiring states to not knowingly conduct or support wrongful acts in cyberspace, including actions that intentionally damage either critical infrastructure or target computer emergency response teams.<sup>34</sup> These are voluntary, legally non-binding commitments that reflect a common standard of acceptable and proscribed behavior, accompanying and expanding on existing legal understandings. They are examples of “Big N norms” that can be described as a form of soft law or codified agreements made by and for states. In other words, norms are what governments say they will do and what they expect others to do. Furthermore, these norms are a result of intentional entrepreneurship by a state. As of now they accompany and expand on existing legal understandings rather than attempt to craft new law, but that does not mean that they may not become legally binding in the future.

While there have been several normative proposals, mostly from China and Russia, on government control over content that directly affects the disinformation discourse, there has yet to emerge such a norm that enjoys broad support from the international community. In lieu of such a norm, Chapter 4 offers options for its development, its legal basis, what it would look like, how it can be promoted, and finally what its risks are.

31 In 2013, the United Nations Group of Governmental Experts in the field of ICT (GGE), the main vehicle within the UN First Committee that deals with international security and disarmament in cyberspace, declared that “international law is applicable and is essential to maintaining peace and stability and promoting an open, secure, peaceful and accessible ICT environment.” United Nations Group of Governmental Experts, “Report of the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/68/98” (June 24, 2013).

32 Louk Faesen, Bianca Torossian, Elliot Mayhew, and Carlo Zensus. *Conflict in Cyberspace: Parsing the Threats and the State of International Order in Cyberspace*. (The Hague: The Hague Centre for Strategic Studies, 2019). <https://hcass.nl/report/conflict-in-cyberspace-parsing-the-threats-and-the-state-of-international-order-in-cyberspace/>.

33 UN GGE, “Report of the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/68/98”; United Nations Group of Governmental Experts, “United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/70/174,” (United Nations: July 25, 2015). <https://undocs.org/A/70/174>

34 Henry Rõigas and Tomáš Minárik, “2015 UN GGE Report: Major Players Recommending Norms of Behaviour, Highlighting Aspects of International Law,” CCDOE, 2015. <https://ccdcoe.org/incyber-arti-cles/2015-un-gge-report-major-players-recommending-norms-of-behaviour-highlighting-aspects-of-international-law/>

## 3.2 Small n norms

The ability of governments to successfully manage the threat of disinformation is not only hampered by the rapid development of digital technologies, but also the dominant role of non-state actors in all shapes and forms (attacker, victim, or carrier of attacks), as well as their unclear relationships with governments. While traditionally all questions related to international peace and security occur within the remit of states and the UN First Committee, in reality governments only constitute one of three stakeholder groups. Cyberspace is largely run by the private sector, which owns and runs most of its digital and physical assets, and the civil society, which is largely responsible for coding and running the global Internet functions.

Those non-state actors also establish norms. In fact, more generally, some of the strongest norms to date, such as the one against slavery, were pioneered by civil society movements. In contrast to governmental norms, which can be fairly abstract and general statements, industry and civil society norms may be much more practical. These are called “small n norms”. They are not a form of soft law that is establishing responsible state behavior, but instead are more practical tested standards and best current practices (BCPs) that are predominantly established by non-state actors.

Small-n norms have been prominent throughout the history of the Internet. For instance, the widespread adoption of the Internet is often traced back to the widespread adoption of two technical standards: TCP/IP and DNS. These standards still continue to be run by a large number of many “informally” (in the legal sense) agreed best current practices. Furthermore, experts have compared “the request for comments process” that lies at the heart of all Internet protocols and standards to “a classic norm-setting process, with all participation in drafting and implementation completely voluntary.”<sup>35</sup> Just like “Big N Norms”, these technical standards also face implementation issues. This often has to do with the implementation costs, which, even if they are minor, are not adequately contrasted with the benefits. Take for example BCP38 on Source Address Validation, a technical norm that dealt with distributed denial of service attacks (DDoS) on networks, which for a long time lagged behind in implementation and compliance until it was incorporated in the Internet Society’s Mutually Agreed Norm of Routing Security (MANRS). In this case, the formulation of a new norm by MANRS overlapped with the existing BCP38 norm and was clearly a reinforcement rather than a distraction. It can take years for “big N Norms” and “small n norms” to be implemented and commonly adhered to. Often the best path to support the acceptance of existing norms is to agree on new add-ons to reinforce existing ones.<sup>36</sup> Similar moves can be seen in the current policy discussions around big N Norms, where states linked a novel norm on the protection of electoral infrastructure as an enhanced interpretation of the pre-existing GGE norm to protecting critical infrastructure.<sup>37</sup>

While norms are certainly a valuable tool, they can be difficult to implement. One of the challenges of agreeing on norms is that they are sometimes formulated by one set of actors but expected to be executed by another. Further, in liberal democracies norms that touch on the freedom of expression – arguably the most important of all democratic principles – require special sensitivity. This requires that the actor groups, regimes, and initiatives fully recognize each other’s mandate or legitimacy. This is not automatically the case. Government actors

<sup>35</sup> Alexander Klimburg and Virgilio Almeida, “Cyber Peace and Cyber Stability: Taking the Norm Road to Stability,” *IEEE Internet Computing* 23, no. 4 (July-Aug. 2019).

<sup>36</sup> Ibid.

<sup>37</sup> Faesen et al., *From Blurred Lines to Red Lines: How Countermeasures and Norms Shape Hybrid Conflict*.

Often the best path to support the acceptance of existing norms is to agree on new add-ons to reinforce existing ones.

can struggle to accept the legitimacy of large internet platforms, while the same non-state actors are often scornful of the knowledge, intention, or capabilities of government. The private sector involved in countering disinformation is also a widely diverse, ranging from social media platforms to Internet Service Providers (ISPs). Working across the regime complex is therefore primarily a question of accepting mutual legitimacy. Any norm, project or initiative that seeks to have a truly global reach and effect on disinformation must have the support of key actors across the regime complex to succeed. These actors are considered to be legitimate either because of their ability to be representative of their constituents (be it members, citizens, or customers), knowledgeable on the technical details within their field, or the ability to practically effect change.

Social media platforms, faced with growing political and broader societal concerns about disinformation, have developed some small n norms that are aimed to address some of the challenges of information disorder, sometimes of their own initiative and sometimes in partnership with governments or civil society organizations. In Chapter 5, small n norms for the social media platforms against disinformation are extrapolated from their current best practices and practical guidelines from academia in Annex I.

### 3.3 Key Takeaways

A norm is a collective expectation for the proper behavior of actors with a given identity, consisting of four core ingredients: (1) identity (the *who*); (2) propriety (the *how*); (3) behavior (the *what* and *where*); and (4) collective expectations (the *why*).

A norm has a lifecycle. This starts with norm *emergence*, in which the norm entrepreneur, who promotes the norm, has the first-mover advantage of *framing* and *linking* the norm to a context of their interest. In the next phases, the entrepreneur must recognize the different strategies and tools of influence at its disposal that allow the norm to *cascade* (wide acceptance) and become *internalized* (deep acceptance).

It can take years for “big N Norms” - government-to-government norms – and “small n norms” – industry best practices and standards – to complete the norm’s lifecycle. Often the best path to support the acceptance of existing norms is to agree on new add-ons to reinforce existing ones.

In some cases, a norm is formulated by one actor group and is expected to be executed by another. The most successful norms in a multifaceted regime complex, such as the one on countering disinformation, requires norms entrepreneurs to work across the regime complex, which is primarily a question of accepting the mandate and legitimacy of the other actor group involved throughout the norm lifecycle.

# 4 Big N Norm against disinformation set by governments

## 4.1 Introduction by Alexander Klimburg

When cyber norms were first proposed in the United Nations context around 2013, they seemed to be a clever fix. Not only did they represent a compromise between two widely separate diplomatic goals, but also very divergent world views and interpretations of how the Internet should be run. This has not changed, and while as of late politically-agreed norms may seemed to have benefited the Western like-minded position, this does necessarily have to stay this way. Norms against disinformation are therefore a case in point.

The intergovernmental norm-setting process in this field has its origins in 1998 when Russia first tabled a UN General Assembly resolution on *information security*. But the definition of information security implicit in this resolution was and remains rather different than the common technical definition of information security, which only was aimed at protecting the confidentiality, integrity and availability of data. Instead, this political definition of information security is clearly associated with domestic stability, covert influencing and insurrection, and what is sometimes called “regime stability”. Most of the Russian proposals that followed sought to introduce and define terms like “information security”, “information war”, “information weapon” and “cyber terrorism”, and to do this within the context of United Nations First Committee dealing with international peace and security.

Russia has regularly re-submitted this resolution in different guises and since 2015 has the strong backing of China as well as the endorsement of the Shanghai Cooperation Organization behind it. Currently known as the Code of Conduct for information security, it effectively amounts to a push for a treaty on cyberspace.<sup>38</sup> This treaty would not only try to regulate state’s malicious behavior in cyberspace, but also shined a light on what kind of malicious behavior was considered most concerning to its authors. This was overwhelmingly orientated towards “bad content”, in other words content that the governments concerned would otherwise not be able to filter and manage the Internet traffic consumed by its citizens, particularly in regard to social media activity and the activity of the press. As has been remarked on before,<sup>39</sup> while Western governments tended to fixate on the threat of “Cyber

38 United Nations General Assembly, “Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General” (United Nations: January 13, 2015). <https://undocs.org/A/69/723>

39 Alexander Klimburg, *The Darkening Web: The War for Cyberspace* (New York: Penguin Books, 2017)

War” and burned-out critical infrastructure as their worst-possible outcome, for states like China and Russia the threat has arguably always been “Information War” (often portrayed as foreign-originated), with their worst possible outcome being regime-change. As such, it was not surprising that the Code of Conduct sets the stage for a very different management of the Internet itself – greatly strengthening the role of states in Internet governance and moving away from the existing bottom-up form of multistakeholder Internet governance, where the private sector and civil society have been in the lead.

The liberal democratic response – long spearheaded by the United States – was originally to oppose any kind of conversation within the UN context. Firstly, the liberal democracies were and are committed to the idea of a non-state-led Internet. Just like the information environment overall (or indeed some physical domains, like the High Seas) are not under universal government jurisdiction, the Internet is considered to be too large and too important to subordinate to a governmental or intergovernmental agency. Secondly, the liberal democracies considered a treaty-based approach to restricting cyber-related malicious activity as a potentially dangerous mirage: such a treaty would be extremely difficult to verify, and while law-abiding nations would adhere to it, others would likely not. The resulting expectation of dishonesty could encourage an even more escalatory cycle of activity than currently being experienced and in a final evaluation would very likely massively increase instability rather than de-escalation. Not inconsequentially, the treaty approach implied that current international law was fell short (perhaps therefore on other issues as well), which was the opposite position of those advocating for a rules-based international order.

However, a minimum compromise was found between the two diplomatic positions of “new international law” and “no new international law”, and that was the initiation of the United Nations Group of Governmental Experts in the field of ICT (GGE) process under the auspices of the UN First Committee (Disarmament). Although initiated in 2003, it led a largely neglected and less productive existence until its third iteration published its consensus report in 2013. This report includes two breakthrough agreements. Firstly, international law applied to cyberspace. Secondly, it introduced norms of behavior as so-called rules of the road for states.

Norms appeared to be the magic compromise everyone was looking for. Liberal democracies opted for this legally non-binding solution as an alternative to new treaties because it maintained the liberal-democratic status quo with existing international law as a point of departure. States like Russia and China, on the other hand, perceived norms not as an enhanced interpretation of existing international law per se, but as evidence that existing law falls short, and those norms are the first step towards new international law drafted by a new status quo.

As both sides seemed to benefit, the process continued. The fourth report of the GGE, published in 2015, strengthened the norms-based approach. Eleven specific norms were agreed upon, including injunctions such as states should not knowingly conduct or support ICT activity that would intentionally damage critical infrastructure.<sup>40</sup> Each of these norms have been rooted in existing international law, and the most recent GGE report (the sixth report of 2021) continued to provide expanded explanation of existing international law via additional examples that were incorporated under the eleven norms that have since become endorsed by the UN General Assembly.

40 UN GGE, “United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/70/174”

Norms appeared to be the magic compromise everyone was looking for.

Given the relative success that cyber norms have had in establishing a common standard of acceptable and proscribed behavior, it is only fair to seriously consider government-to-government norm development in the field of disinformation. However, unlike the norms previously agreed, these norms would be quite different. Firstly, they are arguable not on the spectrum leading up to Article 51 of the UN Charter, where states are allowed to respond to an armed attack in self-defense. Simply put, disinformation, propaganda and covert influencing can never rise to the equivalent of an armed attack according to existing international law, although some countries would certainly like to change this. Secondly, such norms may provide a considerable boost to those states seeking to replace the existing multistakeholder management of Internet resources with an intergovernmental system. Thirdly, they may prove to be an insurmountable test of liberal democracies core values. For most democracies, limitations on freedom of speech already exist, but these are but are carefully managed and constitute the exception, not the rule.

This chapter offers a compromise norm based on the principle of *nonintervention* and *covert* election interference that would prohibit concerted disinformation campaigns and covert influence operations aimed at undermining democratic processes, while allowing overt support for democratic processes and voices, such as public support civil society organizations, political activists or opposition. This also means that *overt* propaganda disseminated by, for example, Russia Today (or indeed by Voice of America, from the Russian perspective) would not be covered by this norm. Rather than focusing on *content*, the norm seeks to restrain certain state conduct, thereby alleviating some of the most important human rights concerns that are linked to discussions curbing online freedom of speech. The application of sovereignty and nonintervention to cyberspace and the wider information environment is still developing and lacking overall consensus among states – even among the like-minded liberal democracies. At the same time, the Covid-19 *infodemic* showed that states are calling out other forms of disinformation outside the context of election interference. A norm proposal should therefore be able to react to this changing environment while at the same time be able to ward off similar concerted Russian proposals of ‘total noninterference’ that prioritize sovereignty over universal human rights, and a multilateral Internet governance system over a multistakeholder one.

Furthermore, in order to steer away from the Russian concept of information security, such a norm proposal is only viable if it includes clear human rights safeguards and if it steers away from any governmental discussion as to what qualifies as “good” or “bad” content. A norm entrepreneur would also be advised to not advance such a norm proposal within the existing UN First Committee forums dedicated to cyber norms. It would only muddle the waters between the agreed norms dealing with cybersecurity and the Russian notion of information security that includes discussions on the content – something that liberal democracies have been relatively successful in avoiding thus far.

## 4.2 The basis for a norm against disinformation

Given the emerging success of recent initiatives to propose, formulate, and develop government-led norms on state behavior in cyberspace, many are looking towards disinformation as an additional threat that could be tackled using a government-to-government norm. As of now, disinformation is not explicitly illegal according to international law, nor is there a



government norm that emerged specifically dedicated to the tackling of disinformation. Drawing from previous research, possible avenues are explored for developing a government-to-government norm against disinformation, its potential legal basis, and offer suggestions on how the norm could be formulated and advanced. Finally, the respective risks are taken into consideration when concluding the utility of such a norm.

What could be a viable basis for government-to-government norm against disinformation? To answer this question, state sovereignty is used as a starting point before moving on to nonintervention. Some believe that sovereignty can erect a normative barrier to disinformation. Within the cyber context, France, Germany, Austria, the Czech Republic and the Netherlands, all support the notion of sovereignty as an enforceable rule, albeit with varying degrees as to what would violate the rule.<sup>41</sup> Out of these states, France has been the most outspoken, stating that “any unauthorized penetration by a state into French systems or any production of effects on French territory via a digital vector may constitute, at the least, a breach of sovereignty”.<sup>42</sup> Following the Russian disinformation campaign targeting the elections in 2017, the French Minister of Defense publicly stated that “by targeting the electoral process of a country, one undermines its democratic foundations, its sovereignty.”<sup>43</sup> However, the position that sovereignty is an enforceable rule in cyberspace, rather than a principle of international law, is far from a universal stance among the like-minded group.<sup>44</sup> By contrast, the US, similarly to the UK, holds the view that sovereignty is merely a principle of international law and does not create autonomous and separate legal obligations, but is protected by other established rules of international law, such as the prohibition of the use of force or nonintervention.<sup>45</sup> As such, without going into the legal details of this debate, sovereignty by itself offers a wide scope of protection to states, but suffers from ambiguity and variations of its interpretation as a rule or a principle.<sup>46</sup>

The nonintervention rule, defined under Article 2(4) of the UN Charter, states that “all Members shall refrain in their international relations from the threat or use of force against the territorial integrity or political independence of any state, or in any other manner inconsistent with the Purposes of the United Nations.”<sup>47</sup> Traditional understandings link the prohibition on the use of force to an element of armed force involved, or at least actions resulting in physical injury or damage. Disinformation has generally sought to test the response thresholds of their opponents by steering clear of causing physical harm, and thereby avoiding the use-of-force threshold. States have also been less open about the application of this threshold to disinformation – a form of statecraft not prohibited under international law. They have not and

41 Przemyslaw Roguski, “The Importance of New Statements on Sovereignty in Cyberspace by Austria, the Czech Republic and United States,” *Just Security*, May 11, 2020, <https://www.justsecurity.org/70108/the-importance-of-new-statements-on-sovereignty-in-cyberspace-by-austria-the-czech-republic-and-united-states/>; Michael Schmitt, “The Defense Department’s Measured Take on International Law in Cyberspace,” *Just Security*, March 11, 2020, <https://www.justsecurity.org/69119/the-defense-departments-measured-take-on-international-law-in-cyberspace/>.

42 Ministry of Defense France, *International Law Applied to Operations in Cyberspace*, (Ministry of Defense: 2019), <https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf#page=6>.

43 Jean-Yves Le Drian, interviewed in *Le Journal du Dimanche*, “France Thwarts 24,000 Cyber-Attacks Against Defence Targets,” BBC, January 8, 2017, <https://www.bbc.com/news/world-europe-38546415>.

44 Przemyslaw Roguski, “The Importance of New Statements on Sovereignty in Cyberspace by Austria, the Czech Republic and United States,” *Just Security*, May 11, 2020, <https://www.justsecurity.org/70108/the-importance-of-new-statements-on-sovereignty-in-cyberspace-by-austria-the-czech-republic-and-united-states/>.

45 Ibid.

46 Gary Corn, “Coronavirus Disinformation and the Need for States to Shore Up International Law,” *Lawfare*, April 2, 2020, <https://www.lawfareblog.com/coronavirus-disinformation-and-need-states-shore-international-law>.

47 United Nations, “Charter of the United Nations,” August 10, 2015, <https://www.un.org/en/charter-united-nations/>.

The position that sovereignty is an enforceable rule in cyberspace, rather than a principle of international law, is far from a universal stance among the like-minded group

are unlikely to deem it as a use of force, largely since doing so would mean that they would inadvertently agree with the Russian and Chinese interpretations of use of force that includes psychological and media warfare.<sup>48</sup>

Nonintervention in the internal affairs of other states is, however, well-established within customary international law: it allows states to safeguard their sovereignty and independence. Moreover, its application to cyberspace has been established and reinforced by many states.<sup>49</sup> However, like the use-of-force prohibition, the nonintervention rule is considered to be of limited scope. Fundamentally, it prohibits the use of *coercive* measures to overcome the free will of a targeted state with respect to matters that fall within that state's core, independent sovereign prerogatives.<sup>50</sup> But, as noted by scholars such as Gary Corn, "the concepts of coercion and "domaine réservé" — the bundle of sovereign rights protected by the rule — are ill defined."<sup>51</sup> Such ambiguities could be cleared up by states disclosing their official views and interpretations. Thus far, only a handful of states have done so on the application of the nonintervention rule in cyberspace and even less for the information environment. Most of these statements only give a general acknowledgment that the parameters of the rule 'have not yet fully crystallized in international law', with some going beyond also mention specific phenomena such as the manipulation of electoral processes and the COVID-19 *infodemic*.<sup>52</sup> The United Kingdom goes further in its statement, namely that an intervention in the fundamental operation of Parliament or in the stability of the financial system would "surely be a breach of the prohibition on intervention."<sup>53</sup>

Arguably, disinformation campaigns that aim to sow discord, distrust, and societal division do not instantly lead to a conclusion of *coercion* as individuals are free to accept and reject information they come across. Some of the responses of governments, however, can provide guidance to the clarification of the coercion element. By linking Russian disinformation in 2016 to fraud and deceit, US Special Counsel Robert Mueller's indictment of the Russian Internet Research Agency and associated individuals<sup>54</sup> demonstrates that covert deception and disinformation can be just as harmful to sovereign prerogative as more overt coercive

48 Russia's and China's perceptions of information as a weapon consider bad content as critical or dissenting of the regime and thereby as an attack against the state. Taylor Cruz and Paulo Simoes, *EECWS 2019 18th European Conference on Cyber Warfare and Security*, (Academic Conferences and Publishing Limited: 2019), 690.

49 The International Court of Justice (ICJ) has described the principle of non-intervention as "a corollary of every state's right to sovereignty, territorial integrity and political independence," and of the right, as a matter of sovereign equality, of every state to conduct its affairs without outside interference. International Court of Justice, "Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)," 1986, <https://www.icj-cij.org/public/files/case-related/70/070-19860627-JUD-01-00-EN.pdf>.

50 Interventions against the sovereignty and the principle of non-intervention require an element of coercion. This concept can be defined broadly or narrowly, with great consequences for the analysis of the case. Unfortunately, international law says very little about the theory of coercion. A complete analysis of what constitutes coercion within this context of international law is too expansive for this study. For more information about this, see Jens David Ohlin, "Did Russian Cyber Interference in the 2016 Election Violate International Law?" *Texas Law Review* 95 (2017): 1579-1598, <https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2632&context=facpub>; Duncan B. Hollis, "The Influence of War; The War for Influence," *Temple International & Comparative Law Journal* 32, no. 1 (2018), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3155273](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3155273).

51 Corn, "Coronavirus Disinformation and the Need for States to Shore Up International Law."

52 The Netherlands referenced to the principle of non-intervention when it called out Russian disinformation campaigns during the COVID-19 pandemic. The Netherlands, "The Kingdom of the Netherlands' response to the pre-draft report of the OEWG." 2020, <https://front.un-arm.org/wp-content/uploads/2020/04/kingdom-of-the-netherlands-response-pre-draft-oewg.pdf>; Corn, "Coronavirus Disinformation and the Need for States to Shore Up International Law".

53 Attorney General's Office and Jeremy Wright, "Cyber and International Law in the 21st Century", *Government of the United Kingdom*, May 23, 2018, <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.

54 United States Department of Justice, "Case 1:18-cr-00032-DLF," United States District Court of for the District of Columbia, February 16, 2018. <https://www.justice.gov/file/1080281/download>.

The concepts of coercion and "domaine réservé" — the bundle of sovereign rights protected by the rule — are ill defined



measures, if not more so.<sup>55</sup> It also reinforces the fact that election processes are a paradigmatic example of the type of sovereign prerogatives protected by the nonintervention rule, leading some legal experts to assert that Russia's election interference crossed a red line.

## 4.3 Framing and linking a norm against disinformation

The principle of sovereignty offers a good starting point for a norm on disinformation but does not provide a sufficiently strong and widely supported basis on itself to develop a norm to address disinformation. This is largely due to the ongoing debate as to whether sovereignty itself is an enforceable rule or merely a principle of international law. Instead, it appears more fruitful to *link* disinformation to the nonintervention principle as certain states view certain aspects of disinformation as exceeding the *coercion* threshold. In particular, election meddling is one of the few forms of disinformation that appears to reach the *coercion* threshold of the nonintervention principle on the basis of official statements or responses from Western like-minded countries.<sup>56</sup>

Furthermore, a norm proposal should also be *framed* in such a way that it prohibits concerted foreign *covert* disinformation and influence campaigns aimed at undermining democratic processes while allowing the US and its partners to both allow and sanction *overt* tools to influence elections, for instance by supporting the civil society in the targeted country through overt formal means, or through the informal support of one's own civil society. This is an important consideration: research shows that, unlike covert influence favored by authoritarian states, most post-Cold War election interference by the United States has been overt, including open support to civil society and democratic processes and aiding governments in the hopes of supporting their reelection.<sup>57</sup> As such, authoritarian regimes, like Russia, would favor a policy of *total* nonintervention and noninterference in the international affairs of other countries, an issue further explored in the section about risks. The suggestion above would form a compromise of sorts: overt means of any sort, including 'propaganda' by state media actors such as Russia Today (or from a Russian point of view BBC or CNN), would be considered acceptable, as would publicly declared funding of civil society organizations (including, for instance, the US National Endowment of Democracy or the Russian Russkiy Mir Foundation). However, hidden subterfuge including clandestine 'civil society' funding, malinformation, or non-transparent strategic communication would become unacceptable.

55 Gary Corn and Eric Jensen, "The Technicolor Zone of Cyberspace – Part I," *Just Security* May 30, 2018, <https://www.justsecurity.org/57217/technicolor-zone-cyberspace-part/>.

56 Lyle J. Moris, Michael J. Mazarr, Jeffrey W. Hornung, Stephanie Pezard, Anika Binnendijk, and Marta Keep, *Gaining Competitive Advantage in the Gray Zone: Response Options for Coercive Aggression Below the Threshold of Major War*, (Santa Monica: RAND Corporation, 2019), [https://www.rand.org/pubs/research\\_reports/RR2942.html](https://www.rand.org/pubs/research_reports/RR2942.html).

57 David Shimer, *Rigged: America, Russia and 100 Years of Covert Electoral Interference*, (London: William Collins, 2020); Peter Beinart, "The U.S. Needs to Face Up to Its Long History of Election Meddling," *The Atlantic*, July 22, 2018, <https://www.theatlantic.com/ideas/archive/2018/07/the-us-has-a-long-history-of-election-meddling/565538/>; Scott Shane, "Russia Isn't the Only One Meddling in Elections. We Do It, Too," *The New York Times*, February 17, 2018, <https://www.nytimes.com/2018/02/17/sunday-review/russia-isnt-the-only-one-meddling-in-elections-we-do-it-too.html>.

A norm proposal should also be framed in such a way that it prohibits concerted foreign covert disinformation and influence campaigns aimed at undermining democratic processes

## 4.4 Promoting the norm against disinformation

Linking the norm to nonintervention and framing it as covert election interference is obviously just one approach that need not forge a 'final government-to-government norm' to the overarching problem of disinformation. But it may constitute a beginning. Beyond these framing and linking options, government norm entrepreneurs can consider starting with a unilateral ban before acquiring broad support to promote the norm.

Robert Knake suggests that the US government should take unilateral action in order to shape global norms in a similar way to the existing norms against commercial intellectual property theft or political assassinations.<sup>58</sup> He believes that, much like how US Executive Order 12333 banned assassinations of political leaders, it would be expeditious to internalize and socialize the norm within the US intelligence community and thus keep the intelligence community from participating in covert election interference.<sup>59</sup> It would not only allow a first-mover advantage in framing the issue but would also combat the perception that liberal democracies such as the US conduct covert influencing activity.<sup>60</sup> The national intelligence community can be persuaded by linking the value of such a norm to national security interests: "In an era in which election interference tools are not held in a Cold War duopoly but are globally available, creating a strong norm against clandestine interference in democratic processes is in the national security interest of the United States."<sup>61</sup>

The government norm entrepreneur should then use a coalition or alliance as an organizational platform to socialize the norm with partners and lay the groundwork for opening multilateral discussions on election interference and to sanction countries that continue to covertly interfere in elections: "As with the agreement with China on economic espionage, the United States and allies would need to agree to abstain from covert election interference even if they are already not doing so in order to allow the Russian government sufficient cover to present any agreement to its citizens as a triumph for Russia."<sup>62</sup> With a norm broadly supported by a wide coalition, the United States or another norm entrepreneur will be better positioned to punish Russia and other rival states if their disinformation campaigns covertly interfere in democratic processes.

## 4.5 The risks of a disinformation Norm

While the previous section offers possible suggestions for *framing* and *linking* norm proposals against disinformation to covert election interference and the nonintervention principle, it already highlights some of the risks. There are serious doubts within the national security

58 Robert Knake, "Banning Covert Foreign Election Interference," *Council on Foreign Relations*, May 29, 2020. [https://www.cfr.org/report/banning-covert-foreign-election-interference?utm\\_medium=social\\_share&utm\\_source=tw](https://www.cfr.org/report/banning-covert-foreign-election-interference?utm_medium=social_share&utm_source=tw).

59 The Executive Order was revoked as part of the Global War on Terror that relied on targeted killings.

60 For a comment on past limitations on US covert influencing activities and the Smith-Mundt Act, see Alexander Klimburg, *The Darkening Web: The War for Cyberspace* (New York: Penguin Books, 2017).

61 Knake, "Banning Covert Election Interference."

62 Ibid.

Total nonintervention would keep Western democracy promotion, support to civil society, aid to opposition parties, public criticism of the Russian regime at bay.

community that Russia would abide to a non-interference pact.<sup>63</sup> More importantly, authoritarian regimes such as Russia actually favor a policy of total nonintervention and noninterference in the internal affairs of other countries and features an important position in many of its diplomatic proposals. Total nonintervention would keep Western democracy promotion, support to civil society, aid to opposition parties, public criticism of the Russian regime at bay.

The first state to actually take a keen interest in proposing norms on this issue was Russia.<sup>64</sup> From 1998 onward, they submitted numerous resolutions that aimed to promote a state-centric conception of information security (see Table 4).<sup>65</sup> Most recently, in September 2020, Russian President Vladimir Putin even suggested a program of measures to restore bilateral relations with the US, that proposed “To exchange, in a mutually acceptable format, guarantees of non-intervention into internal affairs of each other, including into electoral processes, inter alia, by means of the ICTs and high-tech methods.”<sup>66</sup> This proposal is in line with similar pre-existing Russian proposals. Most prominently among these efforts is the *International Code of Conduct for Information Security*, a recurring codification attempt by a Russian-led coalition, stipulating that states subscribing to the Code pledge to “not use information and communications technologies and other information and communications networks to interfere with the internal affairs of other states or with the aim of undermining their political, economic and social stability.”<sup>67</sup> Simultaneously, the Code of Conduct urges for the “establishment of multilateral, transparent and democratic international Internet governance mechanisms.”<sup>68</sup> For the average reader, this may appear as a harmless diplomatic statement, but for those involved, this is considered to be a threat to the existing multistakeholder-led Internet governance ecosystem, which is run by civil society and industry experts and organizations such as the Internet Corporation for Assigned Names and Numbers (ICANN), the Internet Society (ISOC) and the Internet Engineering Task Force (IETF). There is a serious concern that Internet governance resources would be taken away from these civil society organizations and moved towards a multilateral organization, such as the International Telecommunications Union (ITU). This would signify a turn towards a more state-centric, and less civil society-led, governance model of the Internet to the benefit of authoritarian-minded states that would like nothing more than siege root control over the Internet in order to more effectively crack online content that is critical of its regime.

63 Robert Morgus, “Russia Gains an Upper Hand in the Cyber Norms Debate”, *Council on Foreign Relations*, December 5, 2016. <https://www.cfr.org/blog/russia-gains-upper-hand-cyber-norms-debate>

64 See Table 4: Overview of Russian Proposals in the Context of Information Security

65 Stanislav Budnitsky, “Russia’s great power imaginary and pursuit of digital multipolarity,” *Internet Policy Review* 9, no. 3 (August 2020): 9-10.

66 Russia, “Statement by President of Russia Vladimir Putin on a comprehensive program of measures for restoring the Russia – US cooperation in the filed [sic] of international information security,” September 25, 2020. <http://en.kremlin.ru/events/president/news/64086>.

67 Henry Röigas, “An Updated Draft of the Code of Conduct Distributed in the United Nations – What’s New?” CCDCOE, 2015, <https://ccdcoc.org/incyber-articles/an-updated-draft-of-the-code-of-conduct-distributed-in-the-united-nations-whats-new/>.

68 For the 2015 SCO Code of Conduct, see: United Nations General Assembly, “Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General,” January 13, 2015, [https://eucyberdirect.eu/content\\_knowledge\\_hu/2015-sco-international-code-of-conduct-for-state-behaviour-in-information-security/](https://eucyberdirect.eu/content_knowledge_hu/2015-sco-international-code-of-conduct-for-state-behaviour-in-information-security/).

Table 4: Overview of Russian Proposals in the Context of Information Security

Initiative	Main Actors	Year	Outcome
<i>Developments in the field of information and telecommunications in the context of international security</i>	Russia	1998, re-submitted annually to the UN.	Russia called for states to share their perspectives on information security.
<i>Principles of international information security</i>	Russia	1999	A response to their 1998 call for statements on information security, Russia formulated their advocacy for a state-centric approach to internet security.
<i>The UN Group of Governmental Experts (GGE)</i>	United Nations, 20-25 member states	2004, updated four more times (the fifth time was on the basis of US proposal).	Proposal to establish a GGE for discussing international cybersecurity, including on norms. Key success so far include affirming international law applies in cyberspace and 11 proposed norms.
<i>International Code of Conduct for Information Security</i> <sup>69</sup>	SCO, Russia, China	2011, updated in 2015	The authors of this resolution formulated a set of principles that promote a distinct vision on the duties and rights of the state in the information space.
<i>Concept of a Convention on International Information Security</i> <sup>70</sup>	Russia	2011	A 2011 proposal which further defined the Russian perspective on “information security”
<i>World Conference on International Telecommunications (WCIT)</i> <sup>71</sup>	International Telecommunication Union	2013	While many Western states refused to vote, 89 states (led by Russia and China) voted on promoting a state-centric conception of digital rights.
<i>The UN Open-Ended Working Group (OEWG) in the field of information and telecommunications in the context of international security</i>	United Nations	2018-2021; 2021-2026	An international forum open to all member states on issues including cyber norms and behavior.
<i>A comprehensive program of measures for restoring the Russia – US cooperation in the field of international information security</i> <sup>72</sup>	Russia	2020	A proposal in late 2020 which sees Russia again stress its position of non-interference in the internal affairs of each other.

In parallel, Russia and others have pushed to further solidify total noninterference as part of its *Concept of a Convention on International Information Security*.<sup>73</sup> The Russian definition of information security goes beyond the Western or technical definitions and includes content issues (Textbox 2). This allows authoritarian governments to exert more digital control over its citizens, crack down critical and dissenting voices, and secure their regimes.

69 UN GA, “Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General.”

70 The Embassy of the Russian Federation to the United Kingdom of Great Britain and Northern Ireland, “Concept of a Convention on International Information Security,” November 28, 2011, <https://rusemb.org.uk/policycontact/52>.

71 Grnt Gross, “World telecom conference ends with uneven support,” *PCWorld*, December 15, 2012. <https://www.pcworld.com/article/2020583/world-telecom-conference-ends-with-uneven-support.html/>; Grant Gross, “After WCIT, US lawmakers look for ways to advance Internet freedom,” *Computerworld*, February 5, 2013. <https://www.computerworld.com/article/2494615/after-wcit--us-lawmakers-look-for-ways-to-advance-internet-freedom.html>.

72 Russia, “Statement by President of Russia Vladimir Putin on a comprehensive program of measures for restoring the Russia – US cooperation in the filed [sic] of international information security.”

73 The Embassy of the Russian Federation to the United Kingdom of Great Britain and Northern Ireland, “Concept of a Convention on International Information Security.”

## Textbox 2: Information Security vs. information security

### This is not the Information Security you are looking for...

The Western interpretation of *information security* (or infosec) is mostly used by the technical cybersecurity community that relies on the ISO definition, which states that “The purpose of information security is to protect and preserve the confidentiality, integrity, and availability of information (CIA triad). It may also involve protecting and preserving the authenticity and reliability of information and ensuring that entities can be held accountable.” The Western interpretation only concerns itself with the status of the data from a technical point of view (i.e. the CIA triad) not the content of data. The Russian interpretation of *information security* is defined in much broader terms to encompass content issues. It has been defined as “a state in which personal interests, society, and the government are protected against the threat of destructive actions and other negative actions in the information space.”<sup>74</sup> It can encompass critical or dissenting content that is deemed undesirable by the state.

Western states have therefore repeatedly turned down these Russian-led proposals that heavily favor sovereignty over universal human rights, and government-led or multilateral Internet governance over the current multistakeholder model. More specifically, many saw it as an attempt of “strategic revisionism” against established international human rights law, such as free speech, towards a state-centric conception of digital rights. However, the like-minded liberal democracies are a minority compared to the cybersovereignty-oriented countries, such as Russia and China, and many of the G77 countries. This is reflected in the voting results on the International Telecommunication Regulations (ITRs) during the World Conference on International Telecommunications (WCIT) in December 2012 in Dubai. There, Western government representatives were taken by surprise by a resolution on the extension of the ITU mandate to include Internet governance, introducing language which would facilitate government interference with Internet content. In the end, 89 countries, including Russia, China, Saudi Arabia and Singapore, signed the new International Telecommunication Regulations (ITRs) on the spot, while 55 countries, including the United States and nearly all of its like-minded partners, did not (see Figure 3 for the voting results). By voting, the ITU broke tradition of “adopting by consensus” and thus ensured that the new ITRs would not be universally implemented.<sup>75</sup> As some scholars have pointed out, many developing states also prefer Russia and China’s positions advocating the state-centric control of the Internet, as it allows their own national laws and values to take precedence.<sup>76</sup>

74 The Ministry of Foreign Affairs of the Russian Federation, “Convention On International Information Security,” September 22, 2011, [https://www.mid.ru/en/foreign\\_policy/official\\_documents/-/asset\\_publisher/CptlCkB-6BZ29/content/id/191666](https://www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptlCkB-6BZ29/content/id/191666).

75 Alexander Klimburg, “The Internet Yalta,” *Center for a New American Security*, 2, (February, 2013), <https://www.jstor.org/stable/resrep06186>.

76 James A. Lewis, “Liberty, Equality, Connectivity: Transatlantic Cybersecurity Norms,” *Center for Strategic & International Studies*, February 2014: 3-4.

## WCIT Dubai 2012

### Voting results on the ITRs

- Those in favor
- Those against
- Unregistered

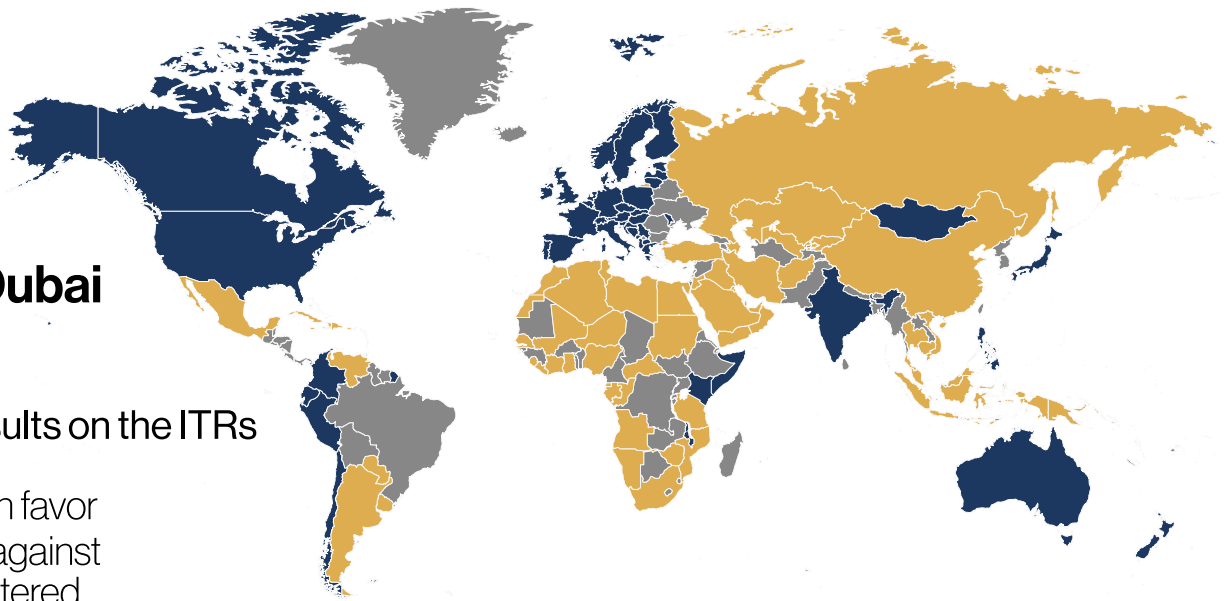


Figure 3: Voting results on the WCIT ITRs.<sup>77</sup>

So, when Western governments publicly argue for a ban on the usage of cyber means to spread disinformation and interfere in the internal affairs of states, they simultaneously need to reject the Russian ready-made noninterference proposals.

Disinformation is a delicate topic to introduce into this mixture. Ironically, Russia is in a better position to advocate the need for binding rules to prohibit noninterference through cyberspace than most Western states. So, when Western governments publicly argue for a ban on the usage of cyber means to spread disinformation and interfere in the internal affairs of states, they simultaneously need to reject the Russian ready-made noninterference proposals. As such, any norm proposal therefore runs the risk of appearing to support Russian proposals or could even be hijacked for this purpose. It is for this reason that Western states insisted on separating the Russian concept of *information security* from the cyber norms discussions within the UN (like the UN GGE and the UN OEWG). They would otherwise move away from technical issues and towards discussions on what content and information should be allowed online, thus opening the door for human rights abusers to argue for sovereign control of information and crack down on dissenting voices via censorship. Russian election interference has placed Moscow's conceptualization of information security front and center in many Western policy discussions, possibly making it harder for them to separate cybersecurity from information security and argue against the need for a more structured and enforceable codification of the noninterference in cyberspace and the larger information environment. Moscow's information warfare is the gift that keeps on giving. A government norm dealing with disinformation would only make it more difficult for Western states to avoid concepts and proposals that are aimed at advancing Russian foreign policy goals.

Similarly, a final issue which makes promoting a norm on disinformation undesirable to many diplomats is how it may undermine existing and ongoing norms discussions in international fora. As mentioned earlier, many Western states have already invested heavily in the current UN OEWG and UN GGE norms processes. Within this context, many of them remain hesitant towards developing additional norms that go beyond the 2015 status quo agreed within the UN GGE. Introducing new norms, especially on disinformation, could lead to current discussions losing focus; undermining or even reformulating the current multilateral norm development efforts in a way that would likely favor Russia.

<sup>77</sup> Derived from Mike Masnick, "Who Signed The ITU WCIT Treaty... And Who Didn't," *Techdirt*, December 14, 2012, <https://www.techdirt.com/articles/20121214/14133321389/who-signed-itu-wcit-treaty-who-didnt.shtml>.



## 4.6 Key Takeaways

A 'big N Norm' proposal against disinformation can be linked to nonintervention and framed around countering covert election interference, after which a norm entrepreneur can consider starting with a unilateral ban before acquiring broad support to promote the norm.

While such a proposal can very well be a viable option that focuses more on *conduct* rather than *content*, thereby alleviating some human rights concerns, many challenges remain. First, the diverging legal interpretations to the application of sovereignty and nonintervention to the wider information environment and what kind of disinformation campaigns constitute a violation of the norm. Second, such a government-to-government norm will make it very difficult for Western states to avoid concepts and proposals that are aimed at advancing Russian foreign policy goals that are advocating for total nonintervention in the internal affairs of other states and a state-centric conceptualization of digital rights and Internet governance. It would therefore require a norm entrepreneur that takes special care to avoid these inadvertent interpretations. Human rights safeguards that protect freedom of speech should be embedded within the norm development process that should preferably not take place within the existing multilateral First Committee forums dedicated to cyber norms as it would risk a move towards the Russian interpretation of information security that includes content issues.

This begs the question: what alternatives exist within a Whole of System approach? Given the significant risks involved, this report suggests another route from the government-to-government "big N Norm" against disinformation. Instead, states should focus on improving the means for content accumulators to better deal with disinformation attempts. One specific area highlighted in recent years is the role of social media platforms. Given their focal role in disseminating disinformation, focusing on platforms appear a promising avenue.

# 5 Small n norms against disinformation set by the industry

## 5.1 Introduction by Frederick Douzet

In December 2020, the European Commission released its Digital Services Act, a new regulatory framework for online platforms designed to better protect users and their fundamental rights online. For the first time, a common set of rules seeks to address online platform obligations and aims to create accountability throughout the entire European Union, aiming to tip the balance of power away from American platforms and into the hands of European policymakers. Clearly, the EU is realizing that the major dependency of European users over American platforms holds major strategic implications: something that becomes especially clear in the field of disinformation and content regulation.

To that extent, the wave of terrorist attacks in the mid-2010s came as a wake-up call for many European countries. The effective use of social media by the Islamic State to disseminate propaganda, control the media agenda, encourage radicalization, recruit soldiers, organize their departure to Syria and raise funds came as a strategic surprise. Not only were European countries unprepared, but they soon realized they had little power over platforms to stop the propagation of Jihadist content. Most platforms initially denied responsibility and resisted cooperation with governments, burnt by the consequences of the Snowden revelations for the trust of their users. They are also concerned with consistency regarding demands coming from governments to avoid being accused of double standards, knowing they are under constant pressure from authoritarian governments to control content and ban specific users. With a business model built around a maximalist view of freedom of expression, content regulation was clearly not in their DNA nor in their plans. Their leaders believed they lacked the skills, the human resources, the tools and the legitimacy to address this fast-growing security threat.

Yet the proliferation of decapitation videos and departures of young people to Syria increased the pressure on social media platforms from both the public and governments. International legal cooperation mechanisms were too slow and too clogged up to address the needs of counter-terrorism agencies. Platforms developed processes of cooperation with governments and civil society through their terms of services to report and take down harmful terrorist content and accounts, with mixed results across countries. Governments also engaged in operations to disrupt Jihadist propaganda, short of being able to remove harmful content or access critical data. But organized violent actors continued to rely on



social networks, with a particularly focal event being the live broadcast on Facebook of the Christchurch mass shooting in New Zealand in 2019. The limitations of the status quo was exposed and caused an international uproar, leading platforms to reluctantly restrict their live-streaming rules while Prime Minister Jacinda Ardern and President Emmanuel Macron launched the “Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online”. Yet, as hard as it might be, reaching consensus at the international level over what constitutes terrorist and violent extremist content is achievable, based on the identification of a specific threat actor to international peace and security<sup>78</sup>. It led to several public-private cooperation mechanisms involving, for example, coordination meetings to exchange best practices, communication procedures to signal particularly viral and harmful content such as Dabiq (ISIS magazine) or online formulas for signaling content that clearly violates the terms of services of the platform. On the other hand, the rise of online disinformation came as a much more complex, yet existential, issue for platforms now directly faced with the consequences of their own power and responsibilities in the aftermath of the 2016 US presidential election.

The 2020 US presidential election appeared as a real stress test for American social media platforms. First, they were concerned about the risk of being leveraged by hostile foreign powers to politically destabilize the American democracy. Since the 2016 election and the subsequent White House accusations against Russia for meddling with the election process, disinformation and digital foreign influence have been identified as a major risk for elections in Western democracies and an important focus of platforms' efforts to identify malicious behavior. Hence, for example, initiatives to identify and take down bots and fake accounts and better control the source of digital ads. With the rise of online conspiratorial networks, such as QAnon, and of organized violent groups taking to social media to organize the storming of the Capitol, platforms were quickly confronted with the harms that real users based in their home country could also perpetrate. This turned out to be a different and trickier challenge for their content moderation teams. After many exhausting months of increasingly violent rhetoric and COVID-19 misinformation, Twitter and Facebook took the extreme step of taking down the US president's accounts.

With the COVID-19 pandemic, the challenge has become global. Conspiracy theorists and anti-vaxxers of all countries, along with a few political leaders, have impeded efforts to fight the propagation of the virus and reduce its mortality. Social media platforms have played a major role in amplifying these theories across the world, to the point that President Joe Biden accused Facebook of “killing people” with misinformation in July 2021.<sup>79</sup>

Social media platforms have come to terms with the fact that they need to monitor and control content more closely in the interest of democracy, as counterintuitive as it might have initially appeared for them. Most of their efforts have focused on identifying deceptive behavior, for instance the diverse techniques of viral deception — whether by automated tools such as bot armies or manual trickery like troll farms — used to enhance the reach, the speed and the scale of their campaign<sup>80</sup>. Targeting conduct, rather than content, to reduce the amplification of disinformation is a more technical and politically less slippery approach. Platforms have also increasingly engaged in identifying and communicating about malicious actors. They

78 United Nations Security Council (UNSC), Res 2129, UN doc S/RES/2129(2013), December 17, 2013, para. 14; et UNSC, Res 2354, UN doc S/RES/2354(2017), May 24, 2017, para. 4 (e).

79 Daniel Politi, “Facebook Pushes Back After Biden Accused It of “Killing People” With Misinformation”, *Slate*, July 18, 2021. <https://slate.com/news-and-politics/2021/07/facebook-pushes-back-biden-killing-people-misinformation.html>

80 Camille François, “Actors, Behaviors, Content: A Disinformation ABC”, *Transatlantic Working Group*, September 20, 2019, [https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC\\_Framework\\_2019\\_Sept\\_2019.pdf](https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf)

Social media platforms have come to terms with the fact that they need to monitor and control content more closely in the interest of democracy, as counterintuitive as it might have initially appeared for them.

have, however, been more reluctant to engage in content control although they have started fighting false information.

Social media companies' action raises multiple legal, ethical, technical and practical questions that get even harder in an international context. Who has the legitimacy to define what disinformation is? Who should enforce the rules? How? What is the appeal process for users? What are the expected results? How do we measure them? Who should have access to the data? What are the risks associated? How do we make sure human and fundamental rights are respected? Is there a democratic oversight? What leverage does the European Union have to defend its own principles and values?

The following section outlines the main measures established by major social media platforms against disinformation, along with the tradeoffs that come with such interventions. Aside from transparency and cooperation mechanisms such as community guidelines (5.2.1) and community reporting and remediation (5.2.8), report offers measures to address deceitful conduct through botnet take down (5.2.2.) and verified information features (5.2.6), along with measures to target harmful content such as factchecking (5.2.3), labelling (5.2.4), verification and labeling of political advertising (5.2.5) and automated control of content under human supervision (5.2.7). All of these measures come with risks associated and require careful implementation associated with transparency measures and performance indicators. This set of measures shows that there are no silver bullets to counter disinformation. To ensure that regulatory proposals help achieve a strengthening of fundamental rights online and a much-needed era of platform accountability, European countries need to fully engage with these nuances to pave the way forward in shaping a digital future that protects users from the online harms.

## 5.2 Small n norms Primer

While government-led norms may be far from ideal in dealing with disinformation, the “small-n norm” avenue focusing on non-state actors appears less risky. Social media platforms, in particular, are at the forefront of the battle against disinformation. In the wake of key focal points, such as the 2016 US elections and the ongoing COVID-19 *infodemic*, these platforms have made commitments and have taken action against disinformation and inauthentic behavior on their platforms – albeit to varying degrees of success and with a lack of coherence between platform efforts. Government pressure turned towards industry to establish stricter standards or impose regulation that aims to minimize the prevalence and impact of disinformation.

Other centralized industry-led approaches exist, albeit primarily dealing with the moderation of more clearly defined and harmful content, including online child abuse and terrorist content (i.e. the Global Internet Forum to Counter Terrorism (GIFCT)). However, Europe (2018) and Australia (2021) have developed Codes of Practice on Disinformation consisting of voluntary commitments by social media platforms, creating a self-regulatory regime. Critics have been vocal of efforts like these, especially the 2018 EU Code of Practice on Disinformation, claiming platforms lack the incentives to meaningfully commit to agreed-upon norms to combat disinformation. As a response, many suggest more stringent coregulatory regimes which would legally mandate compliance to such norms and establish enforcement mechanisms. Both the GIFCT and the EU Code of Practice, and the lessons learned for centralized industry cooperation and the self-regulation approach are analyzed in more detail in Annex II.

Government pressure turned towards industry to establish stricter standards or impose regulation that aims to minimize the prevalence and impact of disinformation.

After a scoping exercise of the companies dealing with disinformation across the Tech Stack (see §5.3) explaining why social media platforms are the most appropriate industry actors for content moderation, this chapter proposes eight standards that can inform an industry charter against disinformation. These standards are developed on the basis of Annex I, which offers a detailed overview of counter-disinformation activities from the largest social media platforms.

## 5.3 Moderation through the Tech Stack

The private sector is arguably the most diverse stakeholder group within the regime complex dealing with disinformation, ranging from social media and news websites that operate closest to the creators and consumers of information, to the domain registrars and Internet Service Providers (ISPs) that operate the technical Internet governance functions. While consumers usually only interact with the user-facing web browsers and platforms, they often forget that powering these is a complex web of interconnected and interdependent Internet services. Together, these form the so-called “Tech Stack” (see Figure 4). Moderating online content and conduct is not only possible at the top levels of the stack, as popularly seen on social media platforms, but also at the lower levels. This section explores the moderation possibilities and risks at each level of the stack to conclude that moderation at the top level is the only credible avenue because it offers the most proportionate response to disinformation.

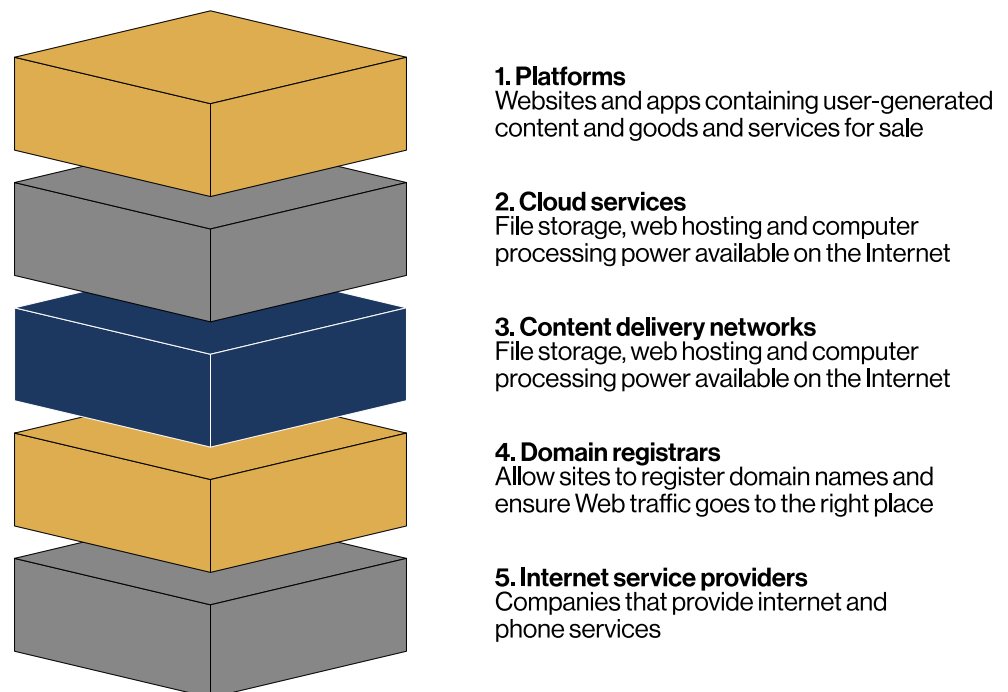


Figure 4: The Tech Stack.<sup>81</sup>

<sup>81</sup> From Geoffrey A. Fowler and Chris Alcantara, “Gatekeepers: These tech firms control what’s allowed online,” The Washington Post, March 24, 2021, <https://www.washingtonpost.com/technology/2021/03/24/online-moderation-tech-stack/>.

## Level 1 – Open Web and Platforms

*E.g. Google, Facebook, Twitter, NYTimes.com, Wikipedia, Uber, etc.*

The level closest to the user consists of the individual websites and social media platforms on which content users interact with what is published. Typically, platforms and websites have policies, community guidelines and terms of use prohibiting illegal or harmful content or conduct (such as hate speech and disinformation) that forms the basis for their moderation. They have a range of measures at their disposal to minimize, demonetize, inoculate, demote or even remove a piece of content, allowing for more proportionate responses than the other levels of the tech stack.<sup>82</sup> However, given the vast amount of content generated by the users on these platforms, it is difficult to always fully enforce these policies. And, as more mainstream platforms ban or counter harmful content, users may migrate to more niche and unmoderated platforms. This layer of the Internet has received the lion's share of attention when it comes to moderation. After all, it is where users create and encounter unsavory content.

## Level 2 – Cloud Services: Cloud computing and hosting providers

*E.g. Squarespace, Wix, WordPress, Shopify, Google Cloud, Microsoft Azure, etc.*

Websites and platforms require storage and computing power in order to run. Today, it is easier and cheaper to pay dedicated companies to host a website and its content.<sup>83</sup> One thereby entrusts these service providers with a substantial amount of power: if they choose to stop providing this service to a website, that website will be forced to find another party to host them. Occasionally there are examples of cloud services stopping their service in response to content, most notably because it violates their terms of service, such as prohibited content (e.g. Child Sexual Abuse Material (CSAM) or piracy) or, in some cases, inciting violence or hate speech. Notable incidents include when, in early 2021, Amazon stopped its service to the app Parler after accusing it of not doing enough to moderate threats of violence, causing Parler to be inactive for several weeks until they found a new provider.<sup>84</sup> The incident demonstrates the type of power that Amazon, which controls the largest share of cloud infrastructure, wields because so many companies rely on it. Finding an alternative provider can be challenging. From this level onward, the breadth of moderation is primarily limited to one tool: suspension or termination of services, which raises obvious proportionality concerns.

## Level 3 – Content Delivery Networks (CDNs)

*E.g. Cloudflare, Akami, Amazon Cloudfront, Microsoft Azure, Epik<sup>85</sup>, Paypal, Apple Pay, etc.*

While not hosting any website content, CDNs help improve the speed, security, and performance of a website, ensuring it remains online, stable, and reliable through balancing traffic across multiple servers.<sup>86</sup> Payment services are related to, and consequently often also lumped into, the category of CDNs. Like CDNs, they ensure the trustworthy, reliable, and easy

<sup>82</sup> Eric Jardine, "Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet," *First Monday* 24, no. 12 (December 2019), DOI: <http://dx.doi.org/10.5210/fm.v24i12.10266>.

<sup>83</sup> Geoffrey A. Fowler and Chris Alcantara, "Gatekeepers: These tech firms control what's allowed online," *The Washington Post*, March 24, 2021, <https://www.washingtonpost.com/technology/2021/03/24/online-moderation-tech-stack/>.

<sup>84</sup> Rachel Lerman, "Parler is back online, more than a month after tangle with Amazon knocked it offline," *The Washington Post*, February 15, 2021, <https://www.washingtonpost.com/technology/2021/02/15/parler-returns-online/>.

<sup>85</sup> Epik offers both CDN and Registrar services. Such a dual role is not uncommon.

<sup>86</sup> Cloudflare, "What is a CDN? | How do CDNs work?" Accessed May 6, 2021, <https://www.cloudflare.com/learning/cdn/what-is-a-cdn/>.

transfer of funds digitally. Since there are only a handful of common payment services which are household names, getting banned by these can be very harmful for content creators hoping to profit from their content.

CDNs, like many others further down the stack, see themselves as neutral service providers for whom content moderation is outside of their *modus operandi*. Notable exceptions do exist, again mainly pertaining to illegal content. Cloudflare's decision to terminate services for the Daily Stormer (2017) and 8Chan (2019) arguably expanded these exceptions to also include incitements to violence and hate speech.<sup>87</sup> In theory, termination does not mean that the content will go away, instead it just becomes slower and more vulnerable to attacks. In practice, termination does, however, leave the content so vulnerable that it can effectively be knocked offline. In fact, in the case of the Daily Stormer, Cloudflare was urged by hackers to terminate their service and "get out of the way" so these hackers could "DDoS this site off the Internet."<sup>88</sup> Cloudflare recognized that DDoS attacks launched by vigilante hackers "subverts any rational concept of justice" and that there was a need for future moderation decisions to be "clear, transparent, consistent and respectful of due process."<sup>89</sup> Two years later, in response to their termination of 8Chan, Cloudflare continues to feel "uncomfortable about their role of content arbiter" and claims that unlike social media platforms, Cloudflare is a "mere conduit for content" that "is not visible to users and therefore cannot be transparent and consistent about their policies".<sup>90</sup> This shows that CDNs such as Cloudflare have very little moderation tools and policies at their disposal and are able to shift their standards relatively flexibly.

## Level 4 – Domain Registrars

*Ex. GoDaddy, Tucows, DreamHost, Epik, etc.*

Domain registrars are companies responsible for registering the domain names of websites. They have been hesitant to moderate, only really removing illegal content, such as a trademark infringement or blacklisting, as deemed by a government order.<sup>91</sup> Major registrars have, however, outlined a "*Framework to Address Abuse*" outlining the types of content that would warrant a response without a court order.<sup>92</sup> This includes CSAM, the illegal distribution of opioids, human trafficking, and specific and credible incitements to violence – all usually covered by a company's terms of service. Notable examples of the latter category include the 2017 and 2018 actions by GoDaddy in the wake of the Charlottesville 'Unite the Right' rally,

87 Cloudflare's termination of the white supremacist website the Daily Stormer in 2017, which caused a brief interruption in the site's operations but quickly resurfaced using a competitor, or the termination of 8chan in 2019 following that forum's role in the terror attacks. Since then, 8chan has rebranded as 8kun and has been bouncing from CDN to CDN, having recently lost their provider again following their roles in the January 6 US Capitol riots. Matthew Prince, "Terminating Service for 8Chan," *Cloudflare*, August 5, 2019, <https://blog.cloudflare.com/terminating-service-for-8chan/>; Matthew Prince, "Why We Terminated Daily Stormer," *Cloudflare*, August 17, 2017, <https://blog.cloudflare.com/why-we-terminated-daily-stormer/>; Tim Elfrink, "A cesspool of hate: U.S. web firm drops 8chan after El Paso shooting," *The Washington Post*, August 5, 2019, <https://www.washingtonpost.com/nation/2019/08/05/chan-dropped-cloudflare-el-paso-shooting-manifesto/>; Kari Paul, Luke Harding and Severin Carrell, "Far-right website 8kun again loses internet service protection following Capitol attack," *The Guardian*, January 15, 2021, <https://www.theguardian.com/technology/2021/jan/15/8kun-8chan-capitol-breach-violence-isp>.

88 Prince, "Why We Terminated Daily Stormer."

89 Ibid.

90 Prince, "Terminating Service for 8Chan."

91 Joan Donovan, *Navigating the Tech Stack: When, Where, and How Should We Moderate Content?*, (Waterloo: Centre for International Governance Innovation, 2019): 17.

92 DNS Abuse Framework, "Framework to Address Abuse," May 29, 2020, [https://dnsabuseframework.org/media/files/2020-05-29\\_DNSAbuseFramework.pdf](https://dnsabuseframework.org/media/files/2020-05-29_DNSAbuseFramework.pdf).

where it terminated several extreme right-wing websites, including the Daily Stormer.<sup>93, 94</sup> Like CDNs and cloud services, domain registrars do not have much experience or willingness to moderate content, and only rely on one mechanism: to terminate their services.

## Level 5 – Internet Service Providers (ISPs)

*Ex. Comcast, AT&T, Verizon, etc.*

ISPs are companies that provide Internet service to users either through broadband or cellular networks. They have substantial power: consider the actions of many authoritarian regimes, who force ISPs to block certain Internet addresses belonging to certain news agencies.<sup>95</sup> ISPs have typically remained absent from discussions on content moderation in liberal democracies, limiting themselves only to blocking illegal content, most CSAM<sup>96</sup> and piracy<sup>97</sup>, through their terms of service. There are only a limited number of ISPs out there, making it very difficult for a banned website to return. The main concern of ISPs is that their limited toolbox (outright bans of whole websites and services) has ‘too wide of a swing’ to be used in a rights-respecting manner. Simply put, the enormity of getting banned by an ISP makes the risk of getting it wrong too big.<sup>98</sup> It might also add grist to the mill of the authoritarian states urging for more government control over Internet governance resources, which includes ISPs.

While moderation is theoretically possible on all levels of the Tech Stack, the underdevelopment, lack of incentives and options, and the distance between the end-user and the lower levels makes content or conduct moderation in the context of disinformation most viable on the top level (i.e. platforms) and too risky at the subsequent levels. For the most part, the bottom levels have one option: termination of services, lacking the much-needed proportionality that is available to platforms. Further down the stack, organizations only deal with moderation through court orders or through their terms of service in case of illegal content or conduct, such as CSAM and piracy, and incitement to violence when focal events caused a massive public outcry for action. This does not include disinformation. Most do not see their role as dealing with the hard questions of content moderation outside of this context and they do not have the policies or mechanisms in place to deal with this in a proportionate and transparent manner. Simply put, platforms on Level 1 are much better positioned to address moderation concerns more proportionally. After all, they are the first level of authority the user turns to in order to deal with content issues in a targeted way that goes beyond simply removal of a piece of content or termination of services further down the stack. This is especially relevant for disinformation, which is not explicitly prohibited under international law, and heavily

93 Fowler and Alcantara, “Gatekeepers.”; Christine Hauser, “GoDaddy Severs Ties With Daily Stormer After Charlottesville Article,” *The New York Times*, August 14, 2017, <https://www.nytimes.com/2017/08/14/us/godaddy-daily-stormer-white-supremacists.html>.

94 Epik, another domain registrar, has been critical of their peers evicting certain websites (such as the 2018 ban of Gab by GoDaddy), going as far as to call out these domain registrars for digital censorship. They now host a number of formerly banned websites, including Gab. Rob Monster, “Why Epik welcomed Gab.com,” *Epik*, November 3, 2018, <https://www.epik.com/blog/why-epik-welcomed-gab-com.html>.

95 An ongoing debate in the West concerning ISPs is the Net Neutrality debate, where advocates claim that ISPs should treat all online content equally – and not slow down, block, or prioritize certain content. Should this equality not apply, ISPs could potentially charge money for users to simply access and popular online service (i.e. you would pay X amount per month for access to Facebook). Wikipedia, “Net neutrality,” Accessed May 6, 2021, [https://en.wikipedia.org/wiki/Net\\_neutrality#:~:text=Network%20neutrality%2C%20most%20commonly%20called,source%20address%2C%20destination%20address%2C%20or](https://en.wikipedia.org/wiki/Net_neutrality#:~:text=Network%20neutrality%2C%20most%20commonly%20called,source%20address%2C%20destination%20address%2C%20or)

96 Chelsea Emery, “Comcast, NetZero agree to block Internet child porn,” *Reuters*, July 29, 2008, <https://www.reuters.com/article/us-comcast-childporn/comcast-netzero-agree-to-block-internet-child-porn-idUSN2935028520080729>.

97 Donovan, *Navigating the Tech Stack*: 18; Fowler and Alcantara, “Gatekeepers.”

98 Corynne McSherry, India McKinney, Jillian C. York, “Content Moderation Is A Losing Battle. Infrastructure Companies Should Refuse to Join the Fight,” *Electronic Frontier Foundation*, April 1, 2021, <https://www.eff.org/deeplinks/2021/04/content-moderation-losing-battle-infrastructure-companies-should-refuse-join-fight>.

For the most part, the bottom levels have one option: termination of services, lacking the much-needed proportionality that is available to platforms.



reliant on context. It also ensures that the civil society and industry-led Internet governance resources remain impartial and unaffected from government demands to regulate 'legal' content that they consider to be harmful to their regime.

In some cases, like the Daily Stormer, the platform or website *itself* may be actively pushing harmful content, so action further down the stack is required. In such exceptional cases, there need to transparent policies in place. The alternative is ad-hoc decisions that leave these companies vulnerable to external – government – pressure to crack down on a certain kind of content.

## 5.4 Small n norms for Social Media Platforms

Moderation appears to be more proportionate and less risky to civil liberties if taken at the top level of the Tech Stack, namely when it is done by social media platforms.

Moderation appears to be more proportionate and less risky to civil liberties if taken at the top level of the Tech Stack, namely when it is done by social media platforms. This section takes a closer look at the respective norms, standards, policies and other relevant initiatives established by the major social platforms in dealing with disinformation. Much of this work is based off the findings presented in Annex I, which includes an overview of the main measures established by major social media platforms (Twitter, Facebook, Microsoft, Google and TikTok) against disinformation. It uses the standards established by the EU Code of Practice on Disinformation (see Annex II) as a starting point and updates it with the most recent company developments and academic assessments to extrapolate best practices and lessons learned that inform the eight standards proposed in this section and guide small-n norm development.

The eight areas for which small-n norms, or standards, are developed for social media platforms include:

- |                         |  |
|-------------------------|--|
| 1. Community Guidelines | 5. Political Advertising               |
| 2. Bot Takedowns        | 6. Verified Information Features       |
| 3. Factchecking         | 7. Algorithms and Automated Moderation |
| 4. Labelling            | 8. Community Reporting and Remediation |

Each standard is described along five categories:

1. *Problem set.* Each standard emerges out of an underlying issue in moderation. Here, these issues are problematized and introduced.
2. *Mitigation by the standard.* Based on historical evidence or research, each standard seeks to mitigate the issue introduced in the problem set.
3. *Requirements.* Each standard requires changes in the behavior of platforms before it is successfully implemented.
4. *Risks.* No measure is perfect and many of the standards proposed have risks associated. These are explored in this section, alongside some solutions to mitigate them.
5. *Transparency/Key performance indicators.* A number of Key Performance Indicators are suggested that incentivize transparency and serve to evaluate the standard and the performance of the platforms.

For more information or concrete examples of the measures undertaken by the social media companies to counter disinformation, see Annex I.



## 5.4.1 Community Guidelines

**Standard:** Community guidelines and Terms of Service should clearly outline social media platforms' policy on disinformation, including definitions of what constitutes a violation and the corresponding consequences..<sup>99</sup>

**Problem set:** While community guidelines and similar platform-to-user agreements are commonplace in all major platforms, many critics are concerned about the comprehensiveness of their coverage of disinformation as a concept. A key example would be Facebook: while it does cover disinformation in several forms within its Community Standards, their Oversight Board has recently ruled that these standards were "inappropriately vague."<sup>100</sup> Other platforms are even more ambiguous. For instance, Google's Community Guidelines nowhere explicitly mention the topics of disinformation, misinformation, false news or inauthentic behavior. On the other hand, platforms like TikTok and Twitter do contain explicit references outlining their stances on disinformation and similar inauthentic behavior. This incoherence between platforms is not only detrimental to a user's understanding of their platform's actions or stances on disinformation but can also harm the platform's response to disinformation as a whole. Community guidelines should contain a clear formulation of what a platform considers as disinformation, misinformation, inauthentic behavior and false news - definitions which are crucial to guide the platform's consistent identification of, and subsequent response to, disinformation.

**Mitigation by the standard:** Establishing a commitment for platforms to clearly include within their community guidelines, or other platform-to-user agreements, clear definitions and explanations of their policies on disinformation should greatly increase both policy transparency and user awareness.

**Requirements:** Perhaps most key is for each platform to clearly outline how they define disinformation. This should allow for both users and researchers to be more aware of what content is not permitted and why, and where the platform sees their role in moderation. Establishing a common requirement for definitions on disinformation would also make strides towards more coherence in definitions. As identified in reviews of the EU Code, there exists very little coherence between platform definitions on key concepts like disinformation, something that is not helped at all by the vagueness and hesitancy which surround the clear definitions that do exist.

**Risks:** The hesitancy of platforms to currently commit to such disclosure in their current community guidelines is not accidental. One of the key struggles with disinformation is simply how to define it. Unlike concepts such as terrorism and child sexual abuse material, what qualifies as disinformation is often very open to interpretation and not easy to capture in a definition. For instance, while opinion or satire is often not *technically* much different than disinformation, in *practice* it is.

<sup>99</sup> The difference between community guidelines and terms of service (ToS) should be made clear to the reader. Terms of Service is the agreement between the user and platform where the user agrees not to take part in a variety of undesired behavior, and is typically presented to the user as a 'terms and conditions' contract that must be signed before using the platform's services. Given the evolving threats facing platforms, what is undesired behavior is typically not clarified within the ToS: instead, they tend to state that users must abide by two elements: the law and the platform's community guidelines. Community guidelines, thus, refer largely to what the expected behavior of users on a platform is. This includes descriptions of what is considered unacceptable behavior or use, accompanied with the actions the platform will take in response. Community guidelines are not typically signed by users: however, it is important that they are clear and transparent since they inform what behavior users can do or expect on a platform.

<sup>100</sup> Oversight Board, "Case Decision 2020-006-FB-FBR," last modified: January 28, 2021, <https://www.oversight-board.com/decision/FB-XWJQBU9A/>.

**Transparency/Key performance indicators:** (1) Clear definitions on concepts including disinformation, misinformation, fake news, and inauthentic behavior in a platform's community guidelines; (2) Clear formulations of what a platform's response will be to any of the defined concepts being present on their platform; (3) An easily accessible appeals process for users who have had content removed, restricted or otherwise affected through platform actions tackling disinformation; (4) Periodical publications providing information on the aforementioned appeals process.

## 5.4.2 Bot Takedowns

**Standard:** Social media platforms should remove malicious bot accounts, botnets, or coordinated inauthentic behavior to ensure that only organic human activity is reflected in various measures of popularity, authority, and influence on social media. They are also encouraged to consider preventative measures that can include authenticity verification measures to help prevent these accounts from being created.

**Problem set:** Malicious actors often rely on botnets for the initial amplification of disinformation campaigns, providing a falsely perceived legitimacy and popularity in support of the falsehood.<sup>101</sup> When bots operate in coordination with other bots, they form botnets which can greatly amplify their effect.<sup>102</sup> There are high levels of coordinated inauthentic behavior on many platforms, much of which may not be reported. Facebook is a good example where often the numbers of its reported bot activity does not completely line up with its self-reported results of bot-removal campaigns. For instance, in 2019 Facebook said 5% of the over 2 billion active accounts were fake<sup>103</sup>, while at the same time it has also admitted to having removed 3 billion fake accounts over a six-month period,<sup>104</sup> meaning that without moderation (and perhaps before 2018 and large-scale bot mitigation strategies) a very large portion of Facebook activity is still driven by bots. Indeed, social media companies may have a vested interest in not driving down bot numbers, given how they also inflate the value of a network. Large platforms have increasingly focused on botnets and have shifted their moderation strategies from supervised approaches aimed at individual accounts and binary label generation

Indeed, social media companies may have a vested interest in not driving down bot numbers, given how they also inflate the value of a network.

101 Maggie Miller, "Social media bots pose threat ahead of 2020," *The Hill*, August 6, 2019, <https://thehill.com/policy/cybersecurity/456282-social-media-bots-pose-threat-ahead-of-2020>; Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini and Filippo Menczer, "The spread of low-credibility content by social bots," *Nature Communications* 9, no. 4787 (2018), DOI: <https://dx.doi.org/10.1038%2Fs41467-018-06930-7>; Ahmed Al-Rawi and Vishal Shukal, "Bots as Active News Promoters: A Digital Analysis of COVID-19 Tweets," *Information* 11 (2020): 461-474. DOI:10.3390/info11100461; Hannes Grassegger and Mikael Krogerus, "Fake news and botnets: how Russia weaponised the web," *The Guardian*, December 2, 2017, <https://www.theguardian.com/technology/2017/dec/02/fake-news-botnets-how-russia-weaponised-the-web-cyber-attack-estonia>.

102 Stieglitz, Brachten, Ross, and Jung identify a variety of different types of bots and botnets. First, they distinguish between benign and malicious bots; with benign bots doing useful services (i.e. aggregating and responding to content), while malicious bots are designed to do harm (i.e. spam, theft of data, spread of malware, and spreading disinformation). Second, they also distinguish between bots who attempt to mimic human behavior to pass off as an authentic user (called 'social bots') versus those who openly portray themselves as bots. Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung, "Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts," *Australasian Conference on Information Systems* (2017), <https://arxiv.org/ftp/arxiv/papers/1710/1710.04044.pdf>.

103 Most fake social media accounts are bots, created by automated programs to post certain kinds of information, which constitutes a violation of Facebook's terms of service. See: Craig Silverman, "Facebook Removed Over 2 Billion Fake Accounts, But The Problem Is Getting Worse," *Buzzfeed*, May 24, 2019, <https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-accounts-afd>; Rob Lever, "Fake Facebook accounts: Never-ending battle against bots," *The Jakarta Post*, May 25, 2019, <https://www.thejakartapost.com/life/2019/05/25/fake-facebook-accounts-never-ending-battle-against-bots.html>.

104 Rob Lever, "Fake Facebook accounts: Never-ending battle against bots," *The Jakarta Post*, May 25, 2019, <https://www.thejakartapost.com/life/2019/05/25/fake-facebook-accounts-never-ending-battle-against-bots.html>.

to techniques that detect suspicious coordination.<sup>105</sup> This has led companies to shift from merely identity verification strategies that characterized inauthentic identity as a status to strategies that understand it as a behavior.<sup>106</sup> Most platforms also prohibit inauthentic behavior in their Terms of Service.

Platforms face both incentives and disincentives to tackle bots. As mentioned above, some point to the fact these bot accounts can boost their traffic and ad revenues,<sup>107</sup> while others warn about the reputational cost or decline in user trust of having many bots on a platform.<sup>108</sup> Since 2016, bot takedowns is one of the areas where platforms have intensified their efforts in the most, albeit primarily focused on bots that attack other users (trolling bots) or bots that hijack coordination tools to render them ineffective (dumping bots), while cheerleading (positive amplification) bots have not received the same amount of attention.<sup>109</sup> A shortage of data available to researchers on platform measures against bots hampers third-party assessments of the effectiveness of the platform performance and should be addressed.

**Mitigation by the standard:** Most platforms have committed to combatting bots and other forms of inauthentic behavior. For instance, the EU Code has a dedicated commitment towards preserving the “Integrity of Services” on platforms. Under this, platforms committed to put into practice clear policies regarding the use and misuse of bots, as well as policies on what constitutes acceptable and unacceptable use of automated systems.<sup>110</sup> Overall, this pillar of the EU Code was generally seen as one of the more effective ones, with a review conducted by the EC concluding that the platforms’ “self-assessment reports demonstrate that platforms have put in place policies to counter the use of manipulative techniques and tactics on their services, including measures to address spammy or inauthentic behavior, fake accounts and malicious, bot-driven activity.”<sup>111</sup> This is further supported by the 2021 EC report *Guidance on Strengthening the Code of Practice on Disinformation*, which focuses primarily on suggesting increasing the cooperation, commitments, and coherence of definitions between platforms.<sup>112</sup> Given this relative success, and the low human rights or free speech risks associated with bot takedowns, an effective standard between platforms that also focuses on cheerleading bots can thus reasonably be achieved. A good start would be requiring more transparency from platforms.

This standard also coincides with the upcoming DSA that requires very large online platforms to “identify, analyze and assess” any risks stemming from the “intentional manipulation of

105 Christian Grimme, Dennis Assenmacher, and Lena Adam, “Changing Perspectives: Is it Sufficient to Detect Social Bots?” *Social Computing and Social Media. User Experience and Behavior* (2018): 445-461, [https://link.springer.com/chapter/10.1007/978-3-319-91521-0\\_32](https://link.springer.com/chapter/10.1007/978-3-319-91521-0_32); Kate Starbird, “Disinformation’s spread: bots, trolls and all of us,” *Nature*, July 24, 2019, <https://www.nature.com/articles/d41586-019-02235-x>; Stefano Cresci, “A Decade of Social Bot Detection,” *Communications of the ACM* 63, no. 10 (September 2020): 72-83, <https://doi.org/10.1145/3409116>.

106 Sarah C. Haan, “Bad Actors: Authenticity, Inauthenticity, Speech, and Capitalism,” *Journal of Constitutional Law* 22, no. 3 (May 2020): 619-686, [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3458795](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458795).

107 Jack Morse, “Why social media companies won’t kill off bots,” *Mashable*, February 7, 2018, <https://mashable.com/2018/02/06/facebook-instagram-twitter-bots/?europe=true#:~:text=Facebook%20isn%27t%20much%20different.&text=The%20more%20they%20engage%20with,contrary%2C%20it%20thrives%20on%20them>.

108 No Author, “Turning a blind eye to bots to protect ad revenue? Think again,” *What’s New in Publishing*, 2019, <https://whatsnewinpublishing.com/turning-a-blind-eye-to-bots-to-protect-ad-revenue-think-again/>.

109 McFaul, Michael. *Securing American Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond* (Stanford: Stanford Cyber Policy Center, June 2019), 60.

110 EC, *Code of Practice on Disinformation*.

111 EC, *Assessment of the Code of Practice on Disinformation*, 9-10.

112 European Commission, *European Commission Guidance on Strengthening the Code of Practice on Disinformation* (Brussels: European Commission, May 26, 2021), 11-13.

their service, including by means of inauthentic use or automated exploitation of the service, with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security.”<sup>113</sup> This would require platforms to release more information on the prevalence of, and their practices against, bots and other forms of inauthentic behavior.

**Requirements:** First, platforms need to clearly identify and define within their policies what kind of inauthentic behavior is and is not permitted. Second, platforms need to take actions to remove inauthentic behavior. This, as evidenced in the self-assessment reports submitted by platforms to the EU Code, primarily takes the form of using technologies like artificial intelligence to automatically detect and block fake accounts. Third, platforms need to be transparent about their efforts. Most EU Code signatories included such transparency within their self-assessment reports, where they disclosed numbers on accounts taken down and posts blocked. Some also publish internal publications which disclose these numbers.<sup>114</sup> However, as mentioned above, this current transparency can be improved upon and expanded to provide third-party researchers with more data and statistics to better evaluate platform efforts.

**Risks:** Overall the risks concerned with shutting down inauthentic accounts and bots are relatively low because the standard focuses on the behavior of platform members rather than the content of their posts. This mitigates possible free speech concerns over content-related moderation.

However, part of the risk with focusing on bot takedowns is that the successes may actually be failures. The fact that platforms like Facebook and Twitter can boast being successful in removing millions of accounts (sometimes each day)<sup>115</sup> is also a reflection on the larger issue that these platforms have with the ease of creating fake accounts. A more meaningful approach could have platforms agreeing to increase the difficulty of creating accounts and adding more advanced user verification features to help prevent these accounts from being created. Yet, focusing on verification posits one of the biggest overall challenges: finding appropriate balance between protecting anonymity, which offers dissenting voices in authoritarian regimes a degree of protection, and enforcing consequences for the abusive behavior, such as disinformation.<sup>116</sup> Decreasing anonymity may result in altered power dynamics between the government and its citizen, “possibly leading to a more inclusive online environment” but also “setting the stage for governments and dominant institutions to even more

113 European Commission, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, (Brussels: European Commission, December 15, 2020). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX-52020PC0825&from=en>.

114 See: Twitter, “Twitter Transparency Center,” Accessed May 11, 2021, <https://transparency.twitter.com/>; Twitter Safety, “Disclosing new data to our archive of information operations,” *Twitter*, September 20, 2019, [https://blog.twitter.com/en\\_us/topics/company/2019/info-ops-disclosure-data-september-2019.html](https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html); Facebook, “Community Standards Enforcement Report,” Last modified February 2021, <https://transparency.facebook.com/community-standards-enforcement#fake-accounts>.

115 Robert K. Knake, “At Facebook, One Million Takedowns Per Day Is Evidence of Failure, Not Success,” *Council on Foreign Relations*, February 20, 2020, <https://www.cfr.org/blog/facebook-one-million-takedowns-day-evidence-failure-not-success>.

116 Nicolo Zingales, “Virtues and Perils of Anonymity: Should Intermediaries Bear the Burden?” *Journal of Intellectual Property, Information Technology and E-Commerce Law* 155 (2014), <https://www.jipitec.eu/issues/jipitec-5-3-2014/4091>; Robert Bodle, “The Ethics of Online Anonymity or Zuckerberg vs “Moot”,” *ACM SIGCAS Computers and Society* 43, no. 1 (May 2013), <https://doi.org/10.1145/2505414.2505417>.

freely employ surveillance tools to monitor citizens, suppress free speech and shape social debate.”<sup>117</sup>

**Transparency/Key performance indicators (KPIs):** The Assessment of the Code of Practice on Disinformation, published by the European Commission in 2020, included a number of suggestions for KPIs on this topic. These are all feasible and provide a good outline of what KPIs on this topic should look like:<sup>118</sup> (1) Ratio (estimate) of inauthentic accounts/users that remained alive/active; (2) Ratio of all engagement (e.g. posts, likes, comments, shares) inauthentic accounts/users have had with genuine users before being detected and deactivated; (3) Ratio of directly contracted employees tasked with identifying/deactivating disinformation content as percentage of total number of staff; (4) Number of fake accounts/fake users deactivated following their report by a genuine user.

### 5.4.3 Factchecking

**Standard:** Social media platforms should have a factchecking process in place, preferably supported by an accredited third party that acts according to independently-established industry standards when labeling disinformation.<sup>119</sup>

**Problem set:** Most of the platforms that use third-party factcheckers often claim it to be highly effective but offer little or no actual data to independently verify such claims.<sup>120</sup> In addition, factchecking often is underfunded and under supported, especially in terms of integration into platform services or the access factcheckers have to relevant information.<sup>121</sup>

**Mitigation by the standard:** Increased transparency is clearly needed to adequately assess the role third-party factcheckers should have in fighting disinformation on platforms, something which initiatives like the EU Code of Practice call for but have not yet successfully delivered. While platforms do not have any industry best practices for the use of factcheckers, there is an International Fact-Checking Network (ICFN) and a code of principles that stipulates standardized practices and accountability for the factchecking industry.<sup>122</sup> Other potential options could include having a third party body, such as the EDMO as proposed by the EC, oversee the collaboration established between platforms and factcheckers and ensure platforms provide factcheckers with sufficient support, integration, and access to adequately do their jobs.<sup>123</sup>

117 Lee Rainie, Jenna Anderson, and Jonathan Albright, “The Future of Free Speech, Trolls, Anonymity and Fake News Online,” *Pew Research Center: Internet and Technology*, March 29, 2017, <https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>.

118 EC, *Assessment of the Code of Practice on Disinformation*, 24.

119 To this extent, the factchecking industry has an International Fact-Checking Network (ICFN) and a code of principles. See Poynter, “The commitments of the code of principles,” *ICFN Code of Principles*, accessed May 6, 2021, <https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles>.

120 Platforms have yet to publicly release data to support the claims. At the same time, studies show that third-party factchecking and labelling can reduce the spread and belief in disinformation: Lee Drutman, “Fact-Checking Misinformation Can Work. But It Might Not Be Enough,” *FiveThirtyEight*, June 3, 2020, <https://fivethirtyeight.com/features/why-twitters-fact-check-of-trump-might-not-be-enough-to-combat-misinformation/>; Man-pui S. Chan, Christopher R. Jones, Kathleen H. Jamieson, and Dolores Albarracín, “Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation,” *Psychological Science* 28, no. 11 (2017), <https://doi.org/10.1177/0956797617714579>; Nathan Walter, Jonathan Coher, and Yasmin Morag, “Fact-Checking: A Meta-Analysis of What Works and for Whom,” *Political Communication* 37, no. 3 (2020), <https://doi.org/10.1080/10584609.2019.1668894>; James Thorne and Andreas Vlachos, “Automated Fact Checking: Task Formulations, Methods and Future Directions,” *Proceedings of the 27th International Conference on Computational Linguistics* (2018), <https://www.aclweb.org/anthology/C18-1283>.

121 EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 19–20.

122 Poynter, “The commitments of the code of principles.”

123 EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 19–20.



**Requirements:** Factchecking is a resource-intensive measure that ideally requires an accepted definition of disinformation and authoritative sources to credibly and objectively discredit a false claim. These elements are often lacking, and the time-intensive processes of fact-checking stands in stark contrast to the quasi-immediacy of contemporary news cycles and information consumption. Additionally, others such as the EC have also called for factchecking to be better integrated and supported by platforms through agreements that are “based on transparent, open and non-discriminatory conditions, and ensure the independence of fact-checkers”. It would also include “fair remuneration to fact-checkers for work used by the platforms, foster cross-border cooperation between fact-checkers, and facilitate the flow of fact-checks across signatories’ services.”<sup>124</sup> It must be noted that while platforms should rely on accredited factcheckers, they are not always readily and widely available in each European language, thus requiring capacity development.

**Risks:** There are three common criticisms against factcheckers: bias, low popular support, and oversight. First, there are concerns about the bias of third-party factcheckers, who have been accused of cherry-picking studies to support their own opinions, which they present as fact.<sup>125</sup> Second, there seems to be a lack of popular support for factchecking on platforms (although it should be noted most of these studies have been conducted in the US where there already is an overall lack of public trust in the media). Third, external factcheckers hired by platforms find that social media platforms retain the last say or authority and can “compel changes to factcheck labels” or “remove misinformation strikes from a page”.<sup>126</sup> Third-party factcheckers may feel that, while social media platforms publicly support and underline the importance of this cooperation, they are too controlling over factchecking.

**Transparency/Key performance indicators:** (1) a company policy stipulating the fact-checking process, (2) the number and share of factchecked content and the average time-frame (3) the share of factchecked content that is flagged as disinformation

#### 5.4.4 Labelling

**Standard:** Social media platforms should create standardized guidelines for a transparent, coherent, and multilevel labeling system dealing with (1) identified disinformation content (potentially with a ranking) (2) sponsored content (including political advertising) and (3) disinformation actors (including machine and human).

**Problem set:** Labelling is a widely-implemented tool that most platforms already use to indicate to a user if a piece of content is perceived as false or misleading, or sponsored. There is, however, a lack of coherency in labelling across platforms, meaning that when disinformation is shared across platforms, the labels are often not. Furthermore, labels are not always clearly visible, and their synergy to other measures, such as Verified Information Features, can be further improved.<sup>127</sup> Currently, most of the labels are attached to pieces of content – be that

<sup>124</sup> Ibid.,

<sup>125</sup> The Editorial Board, “Fact-Checking Facebook’s Fact Checkers,” *Wall Street Journal*, March 5, 2021. <https://www.wsj.com/articles/fact-checking-facebooks-fact-checkers-11614987375>.

<sup>126</sup> Ibid.

<sup>127</sup> All four platforms have a labelling policy. Twitter has gone furthest in enforcing their labelling policies compared to the other platforms. Twitter places interaction limitations to some labels, while Facebook can reduce the distribution of content that is labelled (party) false. YouTube labels have been widely criticized for not visibly depicting its labels, and the video-sharing platform does link its labels to further limitations. The latter point also applies to Microsoft and TikTok. Microsoft’s Bing and Edge uses the NewsGuard labelling system, while TikTok has only recently introduced labels warning for COVID-19 misinformation. (Refer to the ‘Labelling’ subsection in Annex 1).

There are three common criticisms against fact-checkers: bias, low popular support, and oversight.

text or video – rather than the actors behind it, leaving some of the well-known propaganda machines such as RT and Sputnik untouched. Finally, there is a need for more transparency from the platforms on how they label and how effective labels are and to what extent users even see the labels.

**Mitigation by the standard:** While major platforms all label disinformation, none currently share their results or consistent labeling categories with each other – unlike, for instance, cooperation efforts taken to combat terrorist content (see Annex II on GIFCT case study). Coherency across moderation tools can also be improved, for instance by linking the highest-ranking labels to additional content restrictions when new content from the same actor is detected. This can include a link to factchecked information and an advanced warning before or limitations on sharing to slow users down from quickly sharing disinformation, as well as reductions in distribution by algorithms so the labelled content is no longer shared as widely. A widely practiced cybersecurity approach is to “graylist or blacklist” known DNS blocks that distribute malware or even spam – these lists are updated many times a day and are an important part of basic cybersecurity practice for more than a decade.<sup>128</sup> In order for these measures to be proportionate, a multilevel labelling system is required in which only the highest-ranking labels would trigger restrictions. In addition to the existing label regimes for false and sponsored content, another regime can be considered for the producers of disinformation, specifically targeting state-owned or government-controlled or affiliated propaganda. Building on experiences in counterterrorism and cybersecurity, the aim is to have a community resource with which not only the big platforms can profit, but also the smaller social media platforms that normally do not have the ability to enact these measures themselves.

**Requirements:** Coherent labelling across platforms sets the highest requirement as it necessitates a common shared understanding among platforms on what type of content should be addressed with labels, something which changes quite quickly for each platform. For instance, YouTube, Facebook, and Twitter apply labels to state-funded media,<sup>129</sup> while other platforms like Microsoft ban this type of content altogether. This, like the labelling design mentioned before, can be confusing to the user. Ideally, each platform would therefore agree on what type of content falls under each label and improve key technical issues, such as the disappearing labels. Finally, the additional label regime targeting the producers would require a database of disinformation producers and unified standards and language around government-controlled and government-aligned media. Some studies have shown promising results in the mitigating effects it has on misinformation, but only when the labels are noticed and their information is absorbed by the user.<sup>130</sup> This would require exceptional care in the wording of the various categories explaining government involvement to avoid false equivalence between outlets which are editorially independent from governments but receive funding from them, such as the BBC, and outlets closely aligned with government policies, such as Russia Today. If done well, it should inform users of the source of information without unduly harming legitimate journalistic outlets with government ties.<sup>131</sup> The potential of labelling has also been noted by others – e.g. the EC recommended that the Code of Practice “signatories

128 Spamhaus is a leading name in this field. See more at: “Spamhaus.” *Spamhaus*, Accessed on June 6, 2021. <https://www.spamhaus.org/>.

129 Jack Nassetta and Kimberly Gross, “State media warning labels can counteract the effects of foreign misinformation,” Harvard Kennedy School (HKS) Misinformation Review 1, no. 7 (October 2020), DOI: <https://doi.org/10.37016/mr-2020-45>.

130 Ibid.

131 Michael McFaul, *Securing American Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond*, (Stanford: Stanford Cyber Policy Center, June 2019), <https://fsi.stanford.edu/publication/securing-american-elections-prescriptions-enhancing-integrity-and-independence-2020-us>.

There is, however, a lack of coherency in labelling across platforms, meaning that when disinformation is shared across platforms, the labels are often not.



should commit to provide, for all EU languages in which their service is provided, systems for the regular and consistent labelling of content identified as false or misleading and for issuing targeted warnings to users that have interacted with such content.”<sup>132</sup>

**Risks:** There are risks with labelling, most notably the “implied truth effect”, that can be mitigated by additional ‘true’ or ‘unchecked’ labels. Based on the “backfire effect”<sup>133</sup>, the “implied truth effect” suggests that since labelling can never tag 100% of the platform’s content, false content that fails to get tagged can be viewed as truthful and thus is seen as more accurate by consumers than if there were no labels at all.<sup>134</sup> Overall, this risk does not undermine the credibility of labelling as a whole, especially since researchers have examined feasible mitigation methods, such as adding ‘true’ content labels<sup>135</sup> or labelling all content as ‘unchecked’ by default.<sup>136</sup>

**Transparency/Key performance indicators:** (1) a publicly available labelling system for disinformation across platforms, (2) a company policy stipulating the labelling process, (3) the number of labels applied at each level, specified per method and source (if applicable), and their effect on the engagement ratio (i.e. views, clicks, shares), and (4) a database of the producers that were labelled for disinformation.

## 5.4.5 Political Advertising

**Standard:** Social media platforms should take a number of steps to clearly label sponsored content (including political advertising), including requiring verification from the sponsor and having a minimum data reporting requirement on their ad revenue streams. Along these lines, platforms should especially increase their support of current ad repositories to aid researchers. Second, platforms need to increase their oversight over political advertising, as well as limit the targeting capabilities for political advertising.

**Problem set:** Online political advertising is a complex issue that on the one hand has been abused by foreign actors to influence political processes, while also allowing for more political engagement and a more equal level playing field for smaller political parties<sup>137</sup> There are no definitions for issue-based ads, which include political ads, and currently different rules apply

<sup>132</sup> EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 15.

<sup>133</sup> The “backfire effect” is a phenomena when “individuals who receive unwelcome information ... may come to support their original opinion even more strongly.” Brendan Nyhan, and Jason Reifler, “When Corrections Fail: The persistence of political misperceptions,” *Political Behavior* 32 (March 2010): 312, <https://link.springer.com/article/10.1007/s11109-010-9112-2>. However, many more contemporary sources question the relevance of their results. Notably, the 2020 Debunking Handbook writes “The only study to directly examine this notion, however, found no evidence for this effect and instead concluded that a greater number of relevant counterarguments generally leads to greater reduction of misconceptions.” Centre for Climate Change Communication, *The Debunking Handbook* (Washington: The George Mason University, 2020), 10.

<sup>134</sup> Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand, “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings,” *Management Science* 66, no. 11 (November 2020): 4944–4957, <https://doi.org/10.1287/mnsc.2019.3478>.

<sup>135</sup> Ibid.

<sup>136</sup> Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz, and Claire Wardle, “It matters how platforms label manipulated media. Here are 12 principles designers should follow,” *Partnership on AI*, June 9, 2020, <https://www.partnershiponai.org/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/>.

<sup>137</sup> No Author, “Regulate online political ads for greater political integrity”, *Transparency International*, March 10, 2021, <https://www.transparency.org/en/news/regulate-online-political-ads-for-greater-political-integrity>

The combination of the low costs and the ease of targeting to users has made political advertising especially attractive to foreign actors, which is evidenced by actors like the Russian Internet Research Agency (IRA) who especially made use of these features during the 2016 US election cycle.

per country.<sup>138</sup> The forthcoming political ads regulation of the European Commission will help towards much-needed harmonization of online advertisement rules within Europe, especially concerning cross-border adds<sup>139</sup>. Advertising also remains the primary source of revenue for social media platforms. This reliance, however, should not justify platforms to blindly accept any kind of advertising, commercially or politically, without verification. The nature and source of the ads should be made explicit to the platform's users at all times.

The EU Code of Practice on Disinformation had an impact on drawing platform attention to political and issue-based advertising - some platforms imposed limitations and others had already banned political ads altogether. The Code, however, fell short in incentivizing platforms to share relevant and detailed data to be used in assessing the effectiveness of their ad policies. As a result, the European Commission is putting forward new rules on paid online political advertising at the end of 2021 that likely include restrictions of micro-targeting or even prohibitions under certain circumstances.<sup>140</sup> Finally, there exists a need for platforms to actively maintain and support ad repositories that researchers can easily access and use to study the real-time proliferation of advertisements.<sup>141</sup>

The second part seeks to find a way to prevent political advertising tools from being abused by foreign actors to amplify their messaging. This can happen in several manners. First, unlike the organic spread of disinformation, making use of political advertising allows for message amplification limited only by budget. Second, political advertising allows an actor to put content in front of citizens who did not ask to see it or target individuals or groups most vulnerable to their messaging. The combination of the low costs and the ease of targeting to users has made political advertising especially attractive to foreign actors, which is evidenced by actors like the Russian Internet Research Agency (IRA) who especially made use of these features during the 2016 US election cycle.<sup>142</sup>

**Mitigation by the standard:** To address the issue of ad transparency, a focus on increasing transparency overall is needed. More transparency can reduce the negative effects of political advertising as it allows users to distinguish between sponsored and organic content and gives researchers access to data from platforms so they can also play a role in overseeing political ads. One of the ways to do so would be to make the existing ad libraries, or repositories, more usable.

On the second point of the abuse of political advertising tools, actions like limiting the targeting capabilities of political advertising and increasing the verification procedures for those purchasing these ads can be effective. These should also be achievable. As discussed

138 In some European countries, paid political advertising is relatively unrestricted, while in others, political advertising is either allowed only during the pre-election period (for example Italy and Germany) or is completely prohibited (for example in Ireland, France, Belgium, Portugal, Switzerland, and the UK). Jean-François Furnémont and Kevin Deirdre, *Regulation of Political Advertising: a comparative study with reflections on the situation in South-East Europe*, (Council of Europe and the European Union, September 2020), <https://rm.coe.int/study-on-political-advertising-eng-final/1680a0c6e0>

139 The European Commission proposed regulation, titled *Proposal for an initiative on greater transparency in sponsored political content, and other supporting measures*, has three main objectives: "support the functioning of the single market for advertising services; ensure the source and purpose of advertising is known; and combat disinformation and interference in democracy in the EU". See "Political advertising – improving transparency", *European Commission*, [https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Political-advertising-improving-transparency\\_en](https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Political-advertising-improving-transparency_en)

140 Samuel Stolton. "EU executive mulls tougher rules for microtargeting of political ads", *Euractiv*, March 03, 2021. <https://www.euractiv.com/section/digital/news/commission-mulls-tougher-rules-for-microtargeting-of-political-ads/>

141 EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 11; EC, DSA, 9.

142 McFaul, "Securing American Elections," 46.

in a 2019 Stanford publication on electoral interference, “The social media companies have already created voluntary standards for defining political advertisements; they can and should voluntarily choose to limit targeting capabilities for those ads as well.”<sup>143</sup> Through setting more stringent stipulations on what kind of sponsored content is permitted and how it is portrayed to the user, platforms can reduce the likelihood that an ad contains disinformation or is mistakenly identified as organic content by users. However, it is still inevitable that some disinformation can slip through.

Yet, for both ad transparency and abuse of political advertising tools, a more extreme solution exists for platforms: simply banning political advertising on a platform altogether. This has been an option explored by several platforms. Facebook temporarily banned all political advertising from late 2020 to early 2021<sup>144</sup>; and, in 2019, Twitter banned all political advertising on its platform.<sup>145</sup>

**Requirements:** Short of a complete ban on political advertising, there are certain steps that platforms should take to minimize the risks of political or issue-based advertising. The Code of Practice on Disinformation specifically set forth a number of requirements: political ads should be clearly distinguishable from editorial content; signatories should commit to publicly disclosing details behind the ads (such as who sponsored it, much like the US Honest Ads Act that provide transparency by allowing the public to see who bought an online political ad<sup>146</sup>); and signatories should aim to ensure transparency about political advertising.<sup>147</sup> While these targets were not all satisfactorily met, the future may hold more promise, especially with the upcoming EU regulation restricting online political advertising, as well as the proposed DSA that contains provisions stipulating that all online platforms displaying advertisements also clearly display that (a) it is an advertisement; (b) on whose behalf the advertisement is displayed; and (c) meaningful information about why this ad was displayed to someone.<sup>148</sup>

While a 2020 EC review of the Code did note that platforms introduced several actions, such as new policies and political advertising libraries, overall the review found a number of areas which were unsatisfactorily managed, including the identification and public disclosure of issue-based ads, the limited functionalities of libraries made available, and the lack of uniform registration and authorization procedures for political ads.<sup>149</sup> The proposed DSA also contains an article requiring very large online platforms to make their ads publicly available through APIs<sup>150</sup>. This measure has not taken effect yet. A standard addressing this specifically from the perspective of disinformation can add more nuance to ensure coherency and effectiveness.

<sup>143</sup> Ibid.

<sup>144</sup> Mike Isaac, “Facebook Ends Ban on Political Advertising,” *New York Times*, March 3, 2021, <https://www.nytimes.com/2021/03/03/technology/facebook-ends-ban-on-political-advertising.html>.

<sup>145</sup> BBC, “Twitter to ban all political advertising,” *BBC*, October 31, 2019, <https://www.bbc.com/news/world-us-canada-50243306>; Jack Dorsey (@jack), “We’ve made the decision to stop all political advertising on Twitter globally...” *Twitter*, October 30, 2019, [https://twitter.com/jack/status/1189634360472829952?ref\\_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1189634360472829952%7Ctwgr%5E%7Ctwcon%5Es1\\_ref\\_url=https%3A%2F%2Fwww.bbc.com%2Fnews%2Fworld-us-canada-50243306](https://twitter.com/jack/status/1189634360472829952?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1189634360472829952%7Ctwgr%5E%7Ctwcon%5Es1_ref_url=https%3A%2F%2Fwww.bbc.com%2Fnews%2Fworld-us-canada-50243306).

<sup>146</sup> Lau, Tim. “The Honest Ads Act Explained.” *The Brennan Center for Justice*, January 17, 2020. <https://www.brennancenter.org/our-work/research-reports/honest-ads-act-explained>.

<sup>147</sup> EC, *EU Code of Practice on Disinformation*.

<sup>148</sup> EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*.

<sup>149</sup> European Commission, *Assessment of the Code of Practice on Disinformation: Achievements and areas for further improvement*, (Brussels: European Commission, 2020), 9; EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 11.

<sup>150</sup> EC, *DSA*, 63.

In addition, platforms should also, as mentioned earlier, establish more stringent oversight processes for those using political advertising. This includes both being more proactive and demanding in the verification process, as well as limiting the targeting capabilities available. For guidance, platforms could look to cooperate with a variety of civil society initiatives who are working to create standards in this field.<sup>151</sup>

**Risks:** Regulating political advertising, or banning it altogether, receives the most criticism from a rights perspective. Following the bans on political advertising by some platforms, critics were quick to point to the potent impacts on democracy. For instance, online political advertising has enabled entire swathes of the population to become more politically engaged, for better or for worse, and allowed under-funded candidates to use their advertisement budgets more effectively.<sup>152</sup>

**Transparency/Key performance indicators:** The Assessment of the Code of Practice on Disinformation published by the European Commission in 2020, included a number of suggestions for KPIs on this topic. These are all feasible and provide a good outline of what KPIs on this topic should look like.<sup>153</sup> (1) Number of mislabeled political and issue-based ads; (2) Ratio of total turnover of issue-based advertising with revenue lost due to accounts closed down due to breach of issue-based advertising policies; (3) Ratio of number of labelled political advertising against number of political advertising that lost its labelling due to further engagement (e.g. sharing) by platform users per genuine and inauthentic users; (4) Ratio of engagement with labelled political advertising against engagement with political advertising that lost its labelling due to further engagement (e.g. sharing) by platform users per genuine and inauthentic users.

## 5.4.6 Verified Information Features

**Standard:** Social media platforms are encouraged to actively use verified information features, such as links to and pages of factchecked information that debunk disinformation during concerted campaigns or focal points, such as elections or the COVID-19 pandemic. Platforms should also be encouraged to apply such features to other societal issues, such as climate science denial, based on independent disinformation threat assessments. Finally, in exceptional circumstances, a platform should consider using its advertising algorithms to target victims of disinformation with verified information to actively debunk falsehoods.

**Problem set:** During disinformation crises or similar focal points, such as national elections or health emergencies, most Internet users rely on platforms to obtain trustworthy and truthful information. Since platforms themselves are not primarily creators of content, it is useful for them to help spread correct information from authoritative sources.

**Mitigation by the standard:** Many platforms have begun voluntarily spreading authoritative information from international bodies or national governments/agencies to users through a variety of means. This includes features such as information centers/hubs with authoritative information (Facebook, TikTok, Microsoft), information panels (Google) and labels with links

<sup>151</sup> For instance, the Internet Advisory Board has established a number of guidelines for advertising on platforms. See: Internet Advisory Board, "Standards, Guidelines & Best Practices," Accessed on May 11, 2021, [https://www.iab.com/guidelines/?post\\_type=iab\\_guideline](https://www.iab.com/guidelines/?post_type=iab_guideline).

<sup>152</sup> Niam Yaraghi, "Twitter's ban on political advertisements hurts our democracy," *Brookings*, January 8, 2020, <https://www.brookings.edu/blog/techtank/2020/01/08/twitters-ban-on-political-advertisements-hurts-our-democracy/>.

<sup>153</sup> EC, *Assessment of the Code of Practice on Disinformation*, 24.

to authoritative sources (YouTube, TikTok) on videos and other content. These features are often the first line of debunking that users encounter, for instance when a user searches “COVID-19” or a related term on Google, the first result will be an infographic containing authoritative information. Similarly, information centers and hubs are often prominently linked to via labels or prominent posts. While this report introduces the term “Verified Information Feature,” the need for and effect of similar features on user’s access to authoritative information has been reinforced elsewhere. For example, the EC noted in 2021 that signatories of the EU Code of Practice had, for the COVID-19 crisis, “implemented various solutions to provide users with such information and make it visible and easy to access,” and recommended that these signatories should “commit to further develop and apply such specific tools (e.g. information panels, banners, pop-ups, maps and prompts) that prioritize and lead users to authoritative sources on topics of particular public and societal interest or in crisis situations.”<sup>154</sup>

**Requirements:** Verified Information Features differ from platform to platform and may also differ depending on the type of disinformation they are responding to. However, a few common requirements should be in place across these initiatives. First, fact-checked information from authoritative sources should be default. Second, transparency should be front and center. Users should be able to easily isolate where the information used in these features came from and how to access the source information. Third, platforms should ensure these features are simple and easy to understand for all users, including children and elderly. Efforts should especially be made to have these features available in as many different regions or languages as possible and have these features curated by a team local to the regions or knowledgeable in the languages they are available in.

**Risks:** A few risks can be associated with Verified Information Features. Promoting certain content – such as government content – could be seen by some users as contributing to propaganda. As mentioned above, platforms should make sure they are clear about how they select the information displayed in these features. Disclosing, for instance, selection criteria for their sources might be useful. As always, the civil society and academic communities can play a valuable role here in reviewing the sources used by platforms.

In addition, platforms may be tempted to focus heavily on promoting their voluntary creation of these Verified Information Features – as seen, for instance, in platform disclosures made to the EC. This risk especially comes as Verified Information Features are perhaps the anti-disinformation features most immediately encountered by consumers and can be created without changing any of the core processes which may enable the proliferation of disinformation on platforms. Policymakers should remain aware that while Verified Information Features should be encouraged, they are only a small portion of the duties that platforms ought to have to respond to disinformation.

**Transparency/Key performance indicators:** Some suggested KPIs include: (1) The number of visits to/encounters of the feature; (2) The number of clicks when used as a label feature; (3) The number of users using the feature to access additional information; (4) Languages/regions the feature is available in.

<sup>154</sup> EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 14-15.

Two features of algorithms, their personalization and amplification of content, can enable the spread and effect of disinformation.

### 5.4.7 Algorithms and automated moderation

**Standard:** Social media platforms cannot rely on artificial intelligence alone for their moderation but must employ human moderators that are familiar with the local context and language, as well as establish efficient appeal procedures. Platforms need to be transparent about how algorithms work to both suggest and promote content, as well as how they are used in content moderation.

**Problem set:** Most platforms rely on algorithms to recommend and suggest relatable content and advertisements to their users. Two features of algorithms, their personalization and amplification of content, can enable the spread and effect of disinformation.<sup>155</sup> Personalization of content according to the user's views and past behavior can lead to 'filter bubbles' or echo chambers in which users are mainly exposed to content that supports their viewpoint.<sup>156</sup> Algorithms also amplify content as they not only find but also quickly disseminate relatable content to drive engagement. This becomes problematic when algorithms start promoting and spreading false or misleading information because they cannot distinguish it from authoritative content.

Disinformation is a problem that is so pervasive that it cannot be moderated through human moderation alone. Platforms have therefore implemented algorithms to automate moderation, such as content filtering and content removal processes. that, according to the Council of Europe, directly impact freedom of expression and raise rule of law concerns (questions of legality, legitimacy and proportionality).<sup>157</sup> To address these and other human rights concerns, such as the effects on the freedom of assembly and association and the right to free elections, a prominent solution is to increase the algorithmic transparency. Transparency measures can start with platforms providing more information to their users about how algorithms affect the user experience on the platform.<sup>158</sup> Full transparency of the underlying code of algorithms remains a contentious issue, mainly because these companies want to protect their business model (algorithms that stimulate user engagement increase ad revenue) and do not want to reveal to their competitors how their algorithms work.<sup>159</sup>

**Mitigation by the standard:** As noted above, there are two main concerns with algorithms for researchers: first, their usage to suggest and promote content and second, their usage to moderate and remove suspected disinformation. Both usages currently suffer from a lack of transparency. As proven time and again, platforms closely guard their algorithms and their inner workings – often even going as far as to cite them as trade secrets.<sup>160</sup> Yet, transparency is needed. Only by sharing aspects of algorithms can the stakeholders, including the government and civil society, adequately judge the extent that current algorithms in platforms are responsible for both the proliferation of disinformation, as well as the moderation of disinformation. Currently, there is no way to accurately judge the complete impact of algorithmic

<sup>155</sup> Directorate-General for Communications Networks, Content and Technology (European Commission); Open Evidence; and RAND Europe, *Study on media literacy and online empowerment issues raised by algorithm-driven media services* (Luxembourg: Publications Office of the European Union, 2019), 12.

<sup>156</sup> Ibid., 41.

<sup>157</sup> Council of Europe, *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications* (Strasbourg: Council of Europe, 2018).

<sup>158</sup> DG CNECT, Open Evidence and RAND, *Study on media literacy and online empowerment issues raised by algorithm-driven media services*, 59-60.

<sup>159</sup> Ibid., 61-62.

<sup>160</sup> Nazrin, Huseinzade, "Algorithm Transparency: How to Eat the Cake and Have It Too," *European Law Blog*, January 27, 2021, <https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/>.



processes on user experience and rights - a situation which is problematic for policymakers as well as advocates of algorithmic transparency.

**Requirements:** In order to address these concerns, platforms should take several steps. First and foremost, platforms need to ensure effective transparency. Platforms need to make publicly available not only portions or samples of their algorithms and related processes but also statistics about their algorithmic processes - in other words, what is getting taken down and how effective this is. This should then be compared to takedowns of content via traditional human moderation and evaluated whether the tradeoff between speed and quality of moderation is worth it. With the current lack of transparency, such an assessment is not possible for neither academics/civil society nor policymakers to make. Similar statistics should also be made available for the usage of algorithms in recommending and spreading content: how much of this content spread eventually gets flagged as disinformation, how much gets flagged by users, how many users on average see content that is spread by algorithms that is later removed as disinformation, and so on.

Second, platforms need to mitigate the *personalization* and *amplification* effects of algorithms, which are often abused to spread disinformation. This is much easier said than done as these are more inherent characteristics of algorithms rather than unwanted side effects. While improvements to algorithms over time may counter these phenomena, it may still be more promising to increase the recognition of disinformation via algorithmic or machine learning processes, which human content reviewers can then evaluate.

Third, and building off the previous point, platforms need to ensure that if they promote certain content over others (i.e. moderation), then they need to have the capacity to effectively engage in such practices openly and transparently. This entails both the technological capacity to use algorithms or machine learning and the ability to constantly update these processes based on feedback and new innovations, as well as the human resources to do human moderation. This capacity is not something new to platforms. After all, most platforms have been using algorithms for decades, specifically in regards to copyright, terrorism, and illegal speech.<sup>161</sup> Unlike these examples, disinformation is often subtle and very context-dependent. As such, human moderation and overview is still required, and platforms should ensure they have sufficient human resources.

**Risks:** Algorithms are associated with a plethora of risks, which critics do not hesitate to list off. There are several key categories:

- *Human rights.* As noted by the Council of Europe, “Algorithms are widely used for content filtering and content removal processes [...] directly impacting on the freedom of expression and raising rule of law concerns (questions of legality, legitimacy and proportionality).”<sup>162</sup> The majority of these concerns arise from a few issues with algorithms, including their completeness or composition, questions of who decides what algorithms filter out, the amount of personal information these algorithms can compile on large numbers of individuals, and potential bias in the algorithm.<sup>163</sup> These concerns persist and are often brought into mainstream attention by a variety of civil society organizations.

<sup>161</sup> Robert Gorwa, Reuben Binns, and Christian Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” *Big Data and Society* (January-June 2020): 7-10.

<sup>162</sup> Council of Europe, Algorithms and Human Rights, 18.

<sup>163</sup> Ibid.



- *False positives.* AI in disinformation, similar to its usage in other fields, has a troubling tradeoff for platforms. As John Villasenor noted, either social media companies are too expansive in defining disinformation and then risk silencing users posting accurate information or they are too narrow and risk letting disinformation slip through undetected.<sup>164</sup>
- *Bias.* Platforms typically have a global audience and market, which all use the same algorithmic processes. Yet, these processes are primarily developed in the West which can have some biases when covering the Global South, not to mention the optimization of most machine learning processes for the English language.<sup>165</sup> Moreover, other forms of bias by algorithms are also common. Recently, many platforms are beginning to realize that many of their algorithmic processes contain inherent biases based on race and gender; with platforms such as Facebook and Instagram setting up teams and task forces to investigate and respond.<sup>166</sup> Such concerns need to be kept front and center when platforms do engage in moderation based on algorithmic behavior. After all, any content removed accidentally by moderation algorithms due to bias greatly inhibits the public's receptiveness to these technologies and moderation in general.
- *Continual change and evolution.* Platforms and tech companies often boast that they are constantly upgrading and evolving their algorithms.<sup>167</sup> While this is good, it also makes it difficult for external researchers to keep up with all these changes and maintain proper data.
- *Lack of platform incentives to fundamentally change.* More so than the other topics in this subchapter, algorithms are perhaps the area which platforms are least willing to change on. As explored earlier, many even go as far as to view their algorithmic processes as "trade secrets."<sup>168</sup>

**Transparency/Key performance indicators:** (1) Number of content taken down and/or flagged via algorithmic processes for violating disinformation processes; (2) Number of content reported by users for containing disinformation that was approved by algorithms; (3) Amount of users that saw content before it was removed as disinformation via algorithmic processes; (4) Amount of staff working in manual, human review of reported disinformation; (5) Average time it takes algorithms to analyze and remove a piece of disinformation vs human moderation; (6) Frequency of major updates to algorithms. Platforms should also, on a regular basis: (1) Release excerpts of their algorithms to the research community; (2) Provide access to researchers to platform data (i.e. via APIs) to better their understanding of algorithmic processes<sup>169</sup>; (3) Provide users with simple to understand yet informative explanations of how algorithms determine what content they see and how their content is seen by others.

164 John Villasenor, "How to deal with AI-enabled disinformation," *Brookings*, November 23, 2020, <https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation/>.

165 Chinmayi, Arun, "AI and the Global South: Designing for Other Worlds," in *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford: Oxford University Press, 2020). Williams Demetrius, "How Social Media Fact-checking is Inconsistent Across Languages", *TranslateMedia*, June 01, 2021. <https://www.translatemedia.com/translation-blog/how-social-media-fact-checking-is-inconsistent-across-languages/>

166 Deepa Seetharaman and Jeff Horwitz, "Facebook Creates Teams to Study Racial Bias, After Previously Limiting Such Efforts," *The Wall Street Journal*, July 21, 2020. [https://www.wsj.com/articles/facebook-creates-teams-to-study-racial-bias-on-its-platforms-11595362939?mod=hp\\_lista\\_pos1](https://www.wsj.com/articles/facebook-creates-teams-to-study-racial-bias-on-its-platforms-11595362939?mod=hp_lista_pos1).

167 Social Media Today, "Keeping up with the Algorithms," Accessed May 11, 2021, <https://www.socialmediatoday.com/topic/algorithm-updates/>.

168 Huseinzade, "Algorithm Transparency."

169 European Commission, *Tackling online disinformation: a European Approach*, (Brussels: European Commission, 2018), 28.

## 5.4.8 Community Reporting and Remediation

**Standard:** Social media platforms should have a dedicated community reporting mechanism for disinformation, take measures to timely mitigate reported disinformation and to be transparent about their process. Similarly, an platforms should guarantee that individuals have the ability to appeal a decision by the platform.

**Problem set:** All platforms have some way for users to report content but not all have a dedicated option for disinformation. The lack thereof limits the comprehensiveness of the statistics that platforms can collect and share on user-reported disinformation. Some social media platforms state that their internal processes and algorithms are far quicker than relying on community reporting, but many questions remain as to how many user complaints or queries need to be submitted before their automated algorithm identifies it as relevant to be referred to a human operator for content review. Finally, it should be noted that some reports indicate that community reporting receives far less resources.<sup>170</sup>

**Mitigation by the standard:** A dedicated community reporting mechanism for disinformation is one way to decrease the scarcity of up-to-date data that would allow for an assessment of the effectiveness of community reporting vis-à-vis automated mechanisms, while a transparent process constitutes a necessary step towards creating more insight and accountability into the timely response to community feedback. The EC also recently recommended that all platforms should “provide a dedicated functionality for users to flag false and/or misleading information,” noting that these features were not available on all services.<sup>171</sup> However, it did not explore the technical or practical details behind this suggestion in great detail. Instead, it referenced to the DSA act, where Article 17 outlines a requirement for platforms to establish an effective internal complaint-handling system.<sup>172</sup> However, it should again be noted that the DSA does not contain any requirements for this complaint-handling system to have explicit categories on disinformation.

In addition, platforms need to make sure that while they disclose to users when they have removed content on the basis of disinformation, they also incorporate mechanisms for users to dispute these actions if they are perceived to have been committed unfairly. This would take the form of committing to an appeals process or mechanisms for removed content, as well as “appeals transparency”, such as a regular publication outlining appeals and their outcomes. Very few platforms currently provide such offerings.

**Requirements:** Unlike labelling, community reporting does not require the same shared definition of disinformation among platforms, although it should ideally be anchored by a shared common understanding of what type of content the user can flag as disinformation. Community reporting and mediation tools should however be possible in the local language of a user and can be further facilitated by an ombudsman, further explored in Chapter 6.

**Risks:** Just like any other individual moderation tool, community reporting in and of itself will not be sufficient to wholly deal with disinformation. It should not replace automated algorithms or factcheckers but it should function as one of the main tools for facilitating community

<sup>170</sup> See for example the study by the Centre for Countering Digital Hate (CCDH) that showed platforms failed to remove 95% of anti-vaccination misinformation reported to them, and that platforms did not act upon three quarters of the misinformation they reported to them (based on a June 2020 study). It should be noted that these studies had a very limited sample size. Centre for Countering Digital Hate (CCDH), Failure to Act: How Tech Giants Continue to Defy Calls to Rein in Vaccine Misinformation (CCDH: 2020).

<sup>171</sup> EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 15 - 16.

<sup>172</sup> EC, DSA, 53.

A dedicated community reporting mechanism for disinformation is one way to decrease the scarcity of up-to-date data that would allow for an assessment of the effectiveness of community reporting vis-à-vis automated mechanisms.

engagement towards a safe and secure online environment. The main concerns revolve around the low level of participation by the average user in community reporting. Unlike collaborative and community-centered platforms, such as Wikipedia and Reddit, many of the social media platforms do not enjoy the same level of engagement from its users.<sup>173</sup>

**Transparency/Key performance indicators:** (1) a company policy stipulating the community reporting process, including a dedicated reporting mechanism for disinformation and details as to how the mechanism interfaces with automated algorithmic processes (e.g. how many queries does it take before a response is prompted); (2) the number of reports received, actions taken, and the timeframe of the response; (3) the share of user-flagged or algorithm-flagged reports for take-downs or other content restrictions.

---

173 Kaitlyn Tiffany, "Who Would Volunteer to Fact-Check Twitter?" *The Atlantic*, March 3, 2021, <https://www.theatlantic.com/technology/archive/2021/03/twitters-birdwatch-aims-to-crowdsource-fact-checking/618187/>. See also Twitter's Birdwatch project, which is proposed as a community-based approach to countering misinformation. Keith Coleman, "Introducing Birdwatch, a community-based approach to misinformation," *Twitter*, January 25, 2021, [https://blog.twitter.com/en\\_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html](https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html).

## 5.5 Key Takeaways

Overall, small n norms or industry standards against disinformation are less risky than government-to-government big N Norms. They allow for more proportionate measures and raise fewer concerns for human rights violations and for the existing multistakeholder approach to Internet governance. This does not mean these measures do not come with any risks for civil rights and the dominant role of social media platforms as the arbiters of trust. They would require careful implementation associated with transparency measures and performance indicators.

Moderation at the highest levels of the “Tech Stack” through social platforms is most credible, due to their proximity to end users, as well as their ability to take proportionate measures against disinformation. Doing so at lower levels of the Tech Stack is much less proportionate, as these companies mostly rely on one measure: termination of service.

Eight standards against disinformation are proposed as part of a suggested industry charter (see Table 5). These standards are informed by the best practices of the major social media platforms and by academic research. They should be seen as a minimum standard, a starting point from which to build resilience. These eight standards also address some of the shortcomings of the EU Code of Practice on Disinformation and offer more concrete language for the standards, the requirements, risks and the key performance indicators to incentivize transparency and monitor the implementation of the standards by the platforms. Along with other standards proposed elsewhere, they can inform the European Democracy Action plan and the deliberations of the signatories of the Code of Practice as they prepare an update of the Code by the autumn of 2021.

Resources can be directed towards an expert survey to evaluate the formulation of the standards and their feasibility for implementation and effectiveness in countering disinformation. Such a survey could also look at the costs of implementation, the implementation incentives for companies, and the level of support, transparency and monitoring that is currently present and would be required for each standard.

Table 5 Eight proposed standards for an industry charter.

Standard	Description
<b>Community Guidelines</b>	Community guidelines and Terms of Service should clearly outline social media platforms' policy on disinformation, including definitions of what constitutes a violation and the corresponding consequences.
<b>Bot Takedowns</b>	Social media platforms should remove malicious bot accounts, botnets, or coordinated inauthentic behavior to ensure that only organic human activity is reflected in various measures of popularity, authority, and influence on social media. They are also encouraged to consider preventative measures that can include authenticity verification measures to help prevent these accounts from being created.
<b>Factchecking</b>	Social media platforms should have a factchecking process in place, preferably supported by an accredited third party that acts according to independently-established industry standards when labeling disinformation.
<b>Labelling</b>	Social media platforms should create standardized guidelines for a transparent, coherent, and multilevel labeling system dealing with (1) identified disinformation content (potentially with a ranking); (2) sponsored content (including political advertising); and (3) disinformation actors (including machine and human).
<b>Political Advertising</b>	Social media platforms should take a number of steps to clearly label sponsored content (including political advertising), including requiring verification from the sponsor and having a minimum data reporting requirement on their ad revenue streams. Along these lines, platforms should especially increase their support of current ad repositories to aid researchers. Second, platforms need to increase their oversight over political advertising, as well as limit the targeting capabilities for political advertising.
<b>Verified Information Features</b>	Social media platforms are encouraged to actively use verified information features, such as links to and pages of factchecked information that debunk disinformation during concerted campaigns or focal points, such as elections or the COVID-19 pandemic. Platforms should also be encouraged to apply such features to other societal issues, such as climate science denial, based on independent disinformation threat assessments. Finally, in exceptional circumstances, a platform should consider using its advertising algorithms to target victims of disinformation with verified information to actively debunk falsehoods.
<b>Algorithms and automated content moderation</b>	Social media platforms cannot rely on artificial intelligence alone for their moderation but must employ human moderators that are familiar with the local context and language, as well as establish efficient appeal procedures. Platforms need to be transparent about how algorithms work to both suggest and promote content, as well as how they are used in content moderation.
<b>Community Reporting and Remediation</b>	Social media platforms should have a dedicated community reporting mechanism for disinformation, take measures to timely mitigate reported disinformation and be transparent about their process. Similarly, platforms should guarantee that individuals have the ability to appeal a decision by the platform.

# 6 A coregulation model to advance the standards

## 6.1 Introduction by Chris Marsden and Trisha Meyer

Co-regulation ultimately depends on the credible threat of the government to intervene where a self-regulatory scheme fails to achieve its goals. This requires both legislative will and a coherent independent regulator with power to perform its executive function. United Nations Rapporteur Khan also argues for a focus on both legislative powers and effective enforcement: “[s]tate regulation of social media should focus on enforcing transparency, due process rights for users and due diligence on human rights by companies, and on ensuring that the independence and remit of the regulators are clearly defined, guaranteed and limited by law”<sup>174</sup>.

Legislative will is required in order to persuade social media companies of their obligations and the adverse consequences of failure to comply with self-regulatory standards in line with policy pronouncements. This could be seen in 2017-18, when the largest social media companies – known as GAFAM<sup>175</sup> – understood that the European Commission and Parliament were exploring solutions to disinformation in the context of imminent pan-European Parliament elections in May 2019. This gave impetus for social media companies to support the self-regulatory Code of Practice on disinformation adopted under the keen stewardship of the Commission<sup>176</sup>.

However, despite this support from social media companies, many see the EU Code of Practice on disinformation as having numerous issues. Bontridder and Poulet argue that “[t]he Code of Practice on disinformation represents a form of co-regulation that we name ‘ascendant’ since the initiative comes from private actors, the content has been decided by signatories and the execution is marginally controlled by public authorities through the review of the report by the Commission”<sup>177</sup>. Others have criticized the Code for disparity

174 Irene Khan, “Disinformation and freedom of opinion. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan”, *United Nations*, April 13, 2021, para. 91. <https://undocs.org/A/HRC/47/25>.

175 Google-YouTube [Alphabet], Apple, Facebook-Instagram-WhatsApp, Amazon, Microsoft-LinkedIn. In the case of the European code, the signatories include Twitter, Mozilla and TikTok, though these are much smaller organizations.

176 EC, *EU Code of Practice on Disinformation*

177 Noémi Bontridder, and Yves Poulet, “The role of artificial intelligence in disinformation”, *Namur Digital Institute, Faculty of Law*, 2021. <https://researchportal.unamur.be/en/publications/the-role-of-artificial-intelligence-in-disinformation>

between self-reported measures and actual measures; lack of participation of some key platforms (such as WhatsApp) or any independent oversight or even cooperation mechanisms, including with civil society; no independent audit to verify compliance; lack of consequences in case of breach; and lack of protection of fundamental rights, through mechanisms for redress<sup>178</sup>. Thus, in reality, it was self-regulation, not co-regulation<sup>179</sup>.

The lack of a compelling deadline for further moves towards a self-regulatory scheme with sanctions in case of non-compliance has led to a perceived drift in the further adoption of stronger self-regulation. This was accompanied by the uncertainty in GAFAM company headquarters, notably over the fate of the 2020 United States Presidential election. The transfer of power in January 2021 removed that roadblock in the USA, with President Biden in July 2021 accusing Facebook of “killing” people through its failure to remove vaccine misinformation<sup>180</sup>. This American pressure has also coincided with increased legislative will from Europe to move beyond self-regulation: looking forward, impetus towards the 2024 European elections has placed co-regulation firmly back on the political agenda.

The European Commission, announcing its European Democracy Action Plan in December 2020, stated that it “will steer efforts to overhaul the Code of Practice on Disinformation into a co-regulatory framework of obligations and accountability of online platforms, in line with the upcoming Digital Services Act”<sup>181</sup>. The Digital Services Act (DSA) proposal is exactly that<sup>182</sup> – what Bontridder and Pouillet describe as “descending co-regulation”, imposed from above. The Digital Services Act refers to ‘illegal content’, which encompasses part but certainly not all disinformation, with stricter conditions on online advertising, which is again only a subset of disinformation. Since then, the Commission issued guidance to enhance the Code of Practice in Spring 2021 and committed to set up a more robust framework for monitoring its implementation<sup>183</sup>. The DSA will need interpretation in terms of disinformation enforcement, alongside the strengthened 2022 version of the Code of Practice.

This report’s proposal to engage the ‘Whole System’ in combatting disinformation is absolutely necessary. The European Democracy Action Plan is the current best example of such a comprehensive approach, recognizing also the influential role of political parties and the necessity of independent media in particular.

Co-regulation requires a plan for an independent regulatory state-sanctioned backstop when self-regulation fails. In the case of disinformation, the models that can be considered include both electoral and media/audiovisual regulators. The proposed UK Online Safety Bill 2022 is almost as vague as the DSA on disinformation measures but makes clear that Ofcom will be

178 Sounding Board of the Multistakeholder Forum on Disinformation, “The sounding Board’s unanimous final opinion on the co-called Code of Practice”, September 24, 2018. <https://www.ebu.ch/files/live/sites/ebu/files/News/2018/09/Opinion%20of%20the%20Sounding%20Board.pdf>

179 Christopher Marsden, *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*, (Cambridge: Cambridge University Press, 2011) 222; Christopher Marsden, Trisha Meyer, and , Ian Brown, “Platform values and democratic elections: How can the law regulate digital disinformation?”, *Computer Law & Security Review*, 36 (April 2020) 105373. <https://doi.org/10.1016/j.clsr.2019.105373>

180 BBC, “Covid misinformation on Facebook is killing people – Biden”, BBC, July 17, 2021. <https://www.bbc.co.uk/news/world-us-canada-57870778>

181 European Commission, “European Democracy Action Plan: making EU democracies stronger”, press release, December 3, 2020, point 3, [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_20\\_2250](https://ec.europa.eu/commission/presscorner/detail/en/IP_20_2250)

182 EC, *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, art. 35, recitals 68-69.

183 EC, *Guidance on Strengthening the Code of Practice on Disinformation*

The Digital Services Act refers to ‘illegal content’, which encompasses part but certainly not all disinformation, with stricter conditions on online advertising, which is again only a subset of disinformation.



the lead regulator<sup>184</sup>. European Union member states through their coordination mechanism European Regulators Group for Audiovisual Media Services (ERGA) are proposing similar arrangements for audiovisual regulators to regulate disinformation online<sup>185</sup>.

Beyond the need for formal regulators, a coregulation scheme also requires a mechanism that actively promotes and fosters transparency and the sharing of information. This idea is embedded in the 2021 European Commission Guidance on strengthening the Code of Practice, which explains that social media platform “signatories should develop a Transparency Centre where they indicate which policies they adopted to implement the Code’s commitments, how they have been enforced, and display all the data and metrics relevant to the KPIs.” The transparency and conduct of social media platforms in tackling disinformation should remain central, rather than microscopic emphasis on regulating the divergent content types of disinformation. This would set it apart from most current legislative initiatives which tend to focus on content (speech restrictions)<sup>186</sup>.

This report’s recommendations on Information Sharing and Analysis Centers (ISACs) and accountability are pertinent to operationalizing the co-regulation format. There are parallels to David Kaye’s Social Media Council, in particular with regards to monitoring, complaints and ombudsperson. Procedures for making remediation actionable and transparent are crucial, lest the model remain self-regulatory.

The European Commission Guidance also proposes the establishment of a permanent task force chaired by the European Commission, with membership comprising: social media platform signatories, representatives from the European External Action Service, the European Regulators Group for Audiovisual Media Services (ERGA) and European Digital Media Observatory (EDMO). The task force, which would rely on the support of experts, “will help review and adapt the Code in view of technological, societal, market and legislative developments.”

Civil society is included in multiple layers of this report’s co-regulatory “Whole of System” approach – more so than in the European Commission proposal, which remains largely restricted to a role as external expert. In a context where ‘multilateral’ and ‘multistakeholder’ are becoming conflated<sup>187</sup>, not just social media platforms and regulators, but the disinformation community (researchers, fact-checkers, media literacy initiatives) and media need to be engaged in ISAC’s core activities.

Legislative will to promote co-regulation to counteract disinformation is rapidly being supplemented with regulatory monitoring and enforcement institutional reform. The direction towards the 2024 European Parliament elections is clear. The next stage of disinformation co-regulation will need to focus on the effectiveness of these actions.

184 BBC, “Online Safety Bill ‘catastrophic for free speech’”, *BBC*, June 23, 2021. <https://www.bbc.co.uk/news/technology-57569336>

185 European Regulators Group for Audiovisual Media Services, *ERGA Report on disinformation: Assessment of the implementation of the Code of Practice* (ERGA, 2020). <https://erga-online.eu/?p=732>

186 Kalina Botcheva et al., *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression*. Broadband Commission research report on ‘Freedom of Expression and Addressing Disinformation on the Internet’, (International Telecommunication Union and the United Nations Education, Scientific and Cultural Organization, 2020). [https://www.broadbandcommission.org/Documents/working-groups/FoE\\_Disinfo\\_Report.pdf](https://www.broadbandcommission.org/Documents/working-groups/FoE_Disinfo_Report.pdf)

187 See for instance, European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Fostering a European approach to Artificial Intelligence*. (Brussels: European Commission, April 21, 2021), footnote 16.

Legislative will to promote co-regulation to counteract disinformation is rapidly being supplemented with regulatory monitoring and enforcement institutional reform.

## 6.2 Coregulation Primer

A coregulation model is proposed on the European level to advance the standards as part of an industry charter for social media platforms (see Annex III for an overview and comparison of the possible regulatory regimes). This model has two key advantages. First, it avoids a potential situation where platforms could cherry-pick some standards and leave out others. Second, it improves some of the shortcomings of the self-regulatory approach that was previously adopted by the EU Code of Practice mostly concerning transparency, accountability, oversight, and public-private cooperation (see Annex II for the case study on the EU Code of Practice on Disinformation). Finally, an Information Sharing and Analysis Center (ISAC) on Disinformation is proposed to institutionalize the coregulation model and operationalized multistakeholder cooperation to counter disinformation, drawing from lessons learned of the counter-terrorism context (see Annex II: Case Study 1 on the GIFCT) and cybersecurity context (see Annex IV: ISAC's).

It should be noted that the coregulatory approach goes one step further than the self-regulatory direction the European Commission has taken, but the standards for social media platforms and proposals to enhance better accountability and multistakeholder cooperation, most notably the ISAC, remain directly relevant to the upcoming update of the EU Code of Practice on Disinformation as proposed by the European Commission, foreseen early next year.<sup>188</sup> This is especially pertinent once the Code of Practice will be upgraded to a Code of Conduct under article 35 of the Digital Services Act (DSA). The Commission sees the update of the Code as a way for platforms to already prepare for some of the upcoming obligations of the DSA that will be likely be approved and implemented by 2024. Indeed, some components of the Code will be regulated by the DSA, while in other cases the Code goes into more detail or beyond the current proposals of the DSA. The European Commission describes the Code of Conduct process as a coregulatory approach that involves civil society and includes oversight. Oversight in this case does not solely depend on national authorities, but also on a new European Board for Digital Services. The Commission and the Board aim to regularly monitor and evaluate the performance of the very large platforms' performance and update key performance indicators accordingly.

## 6.3 The Regulatory Regimes

To combat disinformation, several regulatory approaches have been considered. Annex III describes three overall approaches based on the work of Chris Marsden, Trisha Meyer and Ian Brown.<sup>189</sup> The "status quo" option entails no regulation and no formal cooperation between the core stakeholder groups – platforms, governments, and civil society – effectively leaving the private sector and the market to deal with the problem of disinformation. The first form of regulation following the "status quo" option is self-regulation, in which industry itself, or in varying degrees of coordination with the government and other stakeholders, creates its own standards to which it will hold itself accountable. This mode appears in three variations: non-audited, audited, and, finally, formal self-regulation in which there is a formal dispute resolution and arbitration mechanism whereby members can be expelled based on non-compliance. Coregulation goes one step further: it is a regulatory application of the multistakeholder

<sup>188</sup> EC, *Guidance on Strengthening the Code of Practice on Disinformation*.

<sup>189</sup> Marsden et al, "Platform values and democratic elections."

Coregulation goes one step further: it is a regulatory application of the multistakeholder approach in which industry sets the standards together with government and civil society. It adds a layer of democratic legitimacy accompanied with independent oversight and enforcement mechanisms.

approach in which industry sets the standards together with government and civil society. It adds a layer of democratic legitimacy accompanied with independent oversight and enforcement mechanisms. Finally, traditional statutory regulation would see the government assuming full responsibility for creating and enforcing regulation. For an overview of each regime and their strengths and weaknesses, see Table 6.

Table 6: Regulatory Regime Options.

Regime	Main features	Examples	Accountability	Transparency	Multistakeholder Cooperation	Enthusiasm of Industry	Protection of Human Rights	Complexity	
Status Quo/ no Regulation	No regulation; single-company led initiatives, corporate social responsibility.	US regulatory environment; platform charters and initiatives (ex. Facebook Oversight Board, community standards).	Low	Low	Low	High	Medium	Low	
Self-regulation	Non-audited	Industry-government coordination, but no sanctions or formalized transparency process (other than potentially self-reporting).	Low	Low	Fairly high	Fairly high	Fairly high	Fairly low	
	Audited	Members are subject to regular, independent audits to judge compliance to agreed-upon criteria.	Potentially the GNI, INHOPE.	Fairly low	Fairly low	Fairly high	Fairly low	Fairly high	Fairly low
	Formal	Can expel non-complying members; dispute resolution/ arbitration on cases; closely supported by existing legislation as the decision body on issues within its mandate.	Video/Video game ratings (e.g. PEGI, IARC).	Fairly high	Fairly high	Fairly high	Fairly low	Fairly high	Fairly low
Coregulation	Industry code approved by Parliament(s) or regulator(s) with statutory powers to supplant	EU Data Protection Directive (1995-2018), Nominet, EURID.	High	High	Fairly high	Fairly low	High	Fairly high	
Statutory or traditional regulation	Formal regulation – tribunal with judicial review	UK Online Harms regime, other forms of traditional, national regulation.	High	High	Fairly poor	Low	High	High	

## 6.4 Why Coregulation?

In order to advance the standards for platforms to counter disinformation, we suggest adopting a similar approach as Marsden and others do, namely: coregulation. It is clear that the libertarian ‘status quo’ approach through which the market would solve the problem is insufficient, both for disinformation as a phenomenon and especially for the European context in general. The self-regulation route has also been tried through the EU Code of Practice on Disinformation, but as Case Study 2 of Annex II shows, this model fell short in effecting meaningful transparency, enforcement, accountability and oversight. And, finally, while traditional or statutory regulation has already been employed by some national governments to establish and enforce standards for the private sector in countering disinformation, such as the French law against the manipulation of information, it is open to criticism concerning government censorship of free press and free speech that can potentially undermine Western-democratic countries’ position vis-à-vis authoritarian states such as Russia. More practically, traditional regulation struggles to keep up with the ever-changing nature of technologies and attack vectors – something that is most efficiently done by directly involving relevant industry actors in the decision-making process to ensure implementation occurs properly.

Coregulation also has a number of unique features which make it especially attractive for both fostering collaboration. Unlike statutory regulation, there is a larger and more prominent role for social media platforms and civil society, yet it improves on self-regulation by including stronger enforcement mechanisms. Coregulation retains the industry involvement in developing standards but adds a “statutory underpinning and legitimacy of parliamentary approval for regulatory systems, together with general principles of good regulation” such as audits, an enforcement mechanism, and an appeal process that are missing in the current self-regulation approach adopted by the EU Code of Practice on Disinformation.<sup>190</sup> In essence, coregulation would see the charter retaining industry involvement in drafting standards but would back these up with a legal framework forcing compliance.

With the connection to the EU legal apparatus, coregulation will uphold a high standard of human and consumer rights. At the same time, it is able to partially sidestep some of the more contentious freedom of expression debates associated with content moderation through leaving the wording and creation of the code up to platforms and other stakeholders. This structure, where a coregulatory body determines the standards but the actors (i.e. social media platforms) themselves are liable for implementation also offers increased flexibility. Flexibility is especially valuable when dealing with different social media platforms and phenomena like disinformation. Coregulation, lastly, also offers more opportunities for users and civil society to offer feedback via mechanisms such as the independent monitoring board. Beyond increasing user rights, this also indicates that successful coregulation would have greater transparency and legitimacy in contrast to more closed off processes like self-regulation.

Finally, it is also valuable to consider the arguments against coregulation. Platforms may still act in ways that are dominated by their business concerns. For instance, industry can decide to not reveal insider knowledge to regulators and instead use its informational edge to push for weaker or imperfect standards that they can then exploit. The model could also contribute to an agency capture, whereby governments pursue the platforms’ agenda rather than that of the public interest. Like self-regulation, business representatives could also free ride on the

<sup>190</sup> Marsden et al, “Platform values and democratic elections,” 14.

In essence, coregulation would see the charter retaining industry involvement in drafting standards but would back these up with a legal framework forcing compliance.

efforts of others, or the bigger companies could have a larger say in setting the standards than their smaller counterparts.

While there are serious concerns for giving industry a great voice in government regulation, coregulation can offer a concrete way in urging platforms to address accountability, transparency, and non-compliance issues that persist in the current self-regulation approach, while leaving sufficient room for each of the platforms to implement the standards in a way that actually works for their unique service.

## 6.5 What a proposed coregulation model could look like

The coregulation model can be visualized as a pyramid in which the regulator sets out a top layer that consists of the overarching regulatory principles and determines the mandate for the coregulatory body consisting of representatives from government, platforms and civil society. The coregulatory body would then translate those principles into an industry charter of standards. Underneath that body, there are the social media companies that would have to implement these standards within their company policies. Next to this pyramid, there would be an oversight board that monitors the degree of implementation and adherence of the standards within the companies, and finally an ISAC which can facilitate public-private threat information exchange and capacity building for the smaller industry partners. This is just one model how platforms can implement the standards through coregulation.

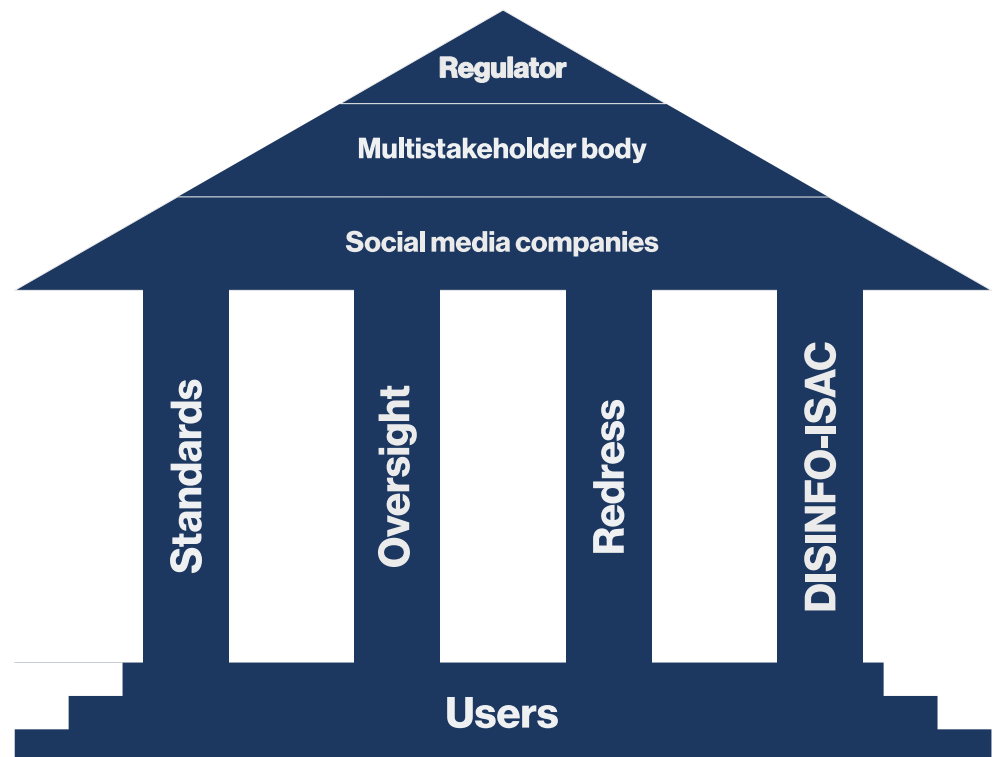


Figure 5: The Coregulation Pyramid and the ISAC Model

### **Regulator: Setting the Principles**

The legislator at the top of the pyramid will set out the overarching thematic pillars and principles for the charter, such as the need for transparency, oversight, remediation, accountability, enforcement, an appeal processes, monitoring, third-party researcher access to data, KPIs, common definitions, the protection of human rights, due process, and the need for human involvement in social media account suspensions, to name just a few.

### **Coregulator: Transposing Principles into Standards**

The multistakeholder body underneath the legislator will be responsible for translating the overarching framework of the regulator and its general principles and objectives into an actionable set of norms and standards, as well as proposing information exchange and capacity building frameworks. Setting the right accountability goals and mechanisms would be part of the multistakeholder body but the actual enforcement would be done through the oversight board.

### **Platforms: Implementing the standards**

Following the adoption of the standards by the coregulator, the platforms must take concrete steps to implement the standards. To some, implementation refers to adoption of the standard, engaging in capacity building efforts, or reaching a granular consensus on the meaning of a standard. While these steps are important prerequisites to implementation, they do not serve to implement the standards themselves. For example, while capacity building is necessary to ensure that platforms can secure themselves and have the bandwidth to engage, one can build capacity without adopting or implementing standards. Rather, implementing a standard involves taking concrete steps to give it force. This might include incorporating the standard into the platform policies or community guidelines and citing the standard when they take far-reaching moderation decisions. Operationalizing a standard in this way also serves to give it a more precise definition.

The regulator and multistakeholder body do not begin their work in a vacuum. Various stakeholder groups have already identified standards. These include the EU and Australian Code of Practice on Disinformation, the Digital Services Act, as well as standards developed by non-state actors, including the platforms themselves. The regulator should therefore first examine how these existing commitments can be better supported and accompanied, where new standards are needed, and how to put these standards into practice. To this end, the proposals in Chapter 5, drafted on the basis of current best practices from social media platforms, can be of use to the multistakeholder body.

### **Oversight Board: Monitoring Platform Performance and Accountability**

An independent oversight board is responsible for monitoring the implementation and effectiveness of the standards. It would do so based on transparency reports of the platforms. The reporting format should be previously agreed by the multistakeholder body and ideally happen in a commonly agreed way, using standardized templates, definitions, and KPIs. The Board should be able to monitor to what extent the platforms have implemented the standards within their community guidelines and company policies and monitor the effectiveness of the standards through the KPIs. Unlike the Facebook Oversight Board, this Board will not be an appeals process that reviews individual moderation decisions of the platforms. Instead, it monitors the overall performance of the platforms in relation to the standards and can issue



policy recommendations to the platforms. It would report to the legislator, who, at some point, could re-convene the multistakeholder body or consider positive inducements or sanctions to improve implementation.

### **Information Sharing and Analysis Center (ISAC): Facilitating Information Exchange and Capacity Building**

Information Sharing and Analysis Centers (ISACs) are traditionally industry-wide non-profit organizations that function as an entrusted entity through which the private sector can share cybersecurity information on threats, vulnerabilities, lessons learned, with each other and with government agencies when appropriate under the law. Its services can range from facilitation of discussions and standard information sharing to operating as an independent and well-staffed intelligence hub. Overall, three unofficial categories can be identified: a country-focused, a sector-specific, and an international collaborative model.

While there are intergovernmental mechanisms for threat information exchange in the context of disinformation, such as the EU Rapid Alert System or the G7 Rapid Response Mechanism, there currently is no formal mechanism or ISAC for social media platforms, nor an industry coordination body that interfaces to EU institutions when it comes to protecting elections from disinformation. For the most part, industry threat information sharing in the context of disinformation occurs on an informal ad-hoc basis. Given the focus of the coregulation model, we suggest to establish a social media industry-wide coalition at the European level. Established by the main platforms, it would allow information exchange between companies, capacity building for smaller members, and coordination with government agencies. It would include a declassification of technical indicators and regular updates on potential threats, with the decision and responsibility still remaining with the individual platforms themselves. A short proposal for such an ISAC is made in Chapter 6.6.

The coregulation model taken in this chapter can be described as a Whole of System approach that pushes for norm advancement in the European context. While it could similarly be applied at the national (Whole of Nation) level, the European approach provides several advantages, mainly having to do with scale, lower transaction costs for the industry, higher leverage for European governments and civil society, and regulatory coherency across European member states. Moreover, doing it at the regional level also builds off the precedents established by the self-regulatory EU Code of Practice on Disinformation, making it a timely addition to the EU Democracy Action Plan that seeks to update the Code.

### **Membership and Responsibilities**

In terms of the membership of the coregulatory scheme, the main challenge will be selecting the relevant civil society members in a legitimate and representative way, while balancing the weight of their voice compared to the platforms. This was also one of the shortcomings in the EU Code of Practice, where the Sounding Board mainly constituted of media representatives and hardly any civil society members from the human rights community, academia or legal experts. The self-regulation model adopted by the Code fell short in facilitating actual multistakeholder participation. For instance, throughout the development of the Code, many non-state stakeholders in the Board criticized the Code for failing to include meaningful actions, KPIs, or commitments. Yet, these concerns were ultimately ignored. A coregulatory regime, with a defined role for civil society in the coregulatory body could counter these previous shortcomings by giving these actors a real voice. Their role and voice would also be

For the most part, industry threat information sharing in the context of disinformation occurs on an informal ad-hoc basis.

strengthened by the inclusion of effective enforcement mechanisms which takes civil society concerns into consideration.

The mere presence of European or national NGOs does not mean that those with the most expertise or those most affected are represented. Without clear selection criteria, most NGOs that participate in the body are hand-picked or self-selecting. Ideally, participation should seek to involve the main actors working in the European regime complex. Working across the European regime complex is, as previously noted, primarily a question of accepting mutual legitimacy. Any norm, standard, or initiative that seeks to have a European reach and effect on disinformation must have the support of key actors across the regime complex to succeed. These actors are considered to be legitimate either because of their ability to be representative of their constituents (be it members, citizens, or customers), knowledgeable on the technical details within their field, or their ability to practically enact change. Accepting any one of these definitions of legitimacy is the equivalent of trusting the verdict of these actors to at least be relevant within the wider discourse of disinformation, thus ensuring trust is present among these different state and non-state actors.

### Accountability

The biggest lesson that can be learned from the EU Code of Practice is accountability. As this report is not alone in noting, the EU Code of Practice fundamentally fell short in this area as this self-regulatory framework was found to “not establish an independent oversight mechanism for monitoring the completeness and impact of the signatories action in tackling disinformation”<sup>191</sup>. Based on this report’s analysis of the GIFCT and the Code, accountability requires four key components: organizational and reporting transparency, meaningful independent oversight, adherence to fundamental rights, and remediation mechanisms.

First, the starting point for accountability is creating organizational and reporting transparency. For a coregulatory model, transparency should be embedded as a key focus initially set by the top-level regulator. This regulator should dictate clear transparency goals, such as access for researchers to datasets and clarity in platform takedowns of content. The coregulatory body then puts these principles into practice. This means platforms should report about their implementation of these standards in their policies and their resulting impact in countering disinformation. Reporting should happen in a commonly agreed way, using standardized templates<sup>192</sup>, definitions<sup>193</sup> and KPIs.<sup>194</sup> KPIs are the measurements chosen in a regulatory scheme to which platforms have to report by. Crafting the right KPIs is of the utmost importance as it sets the incentives for platforms and would thus benefit from a wide range of expert views from all stakeholder groups. These KPIs should aim to overcome the issues of the Code, in which platforms were reluctant in sharing statistics, datasets and insights into the effectiveness of their tools. While the KPIs in a coregulatory regime would still be designed largely by industry (through their position in the coregulatory body), mechanisms such as the legislative component would be much more involved in ensuring that these KPIs are effective

191 Platforms have demonstrated to be reluctant and, occasionally, outright unwilling to provide statistics, datasets, and insights into how the effectiveness and operations of their tools and mechanisms that tackle disinformation. EC, *Assessment of the Code of Practice on Disinformation*, 18.

192 In the transparency reports of the Code of Practice, platforms did not have reporting templates, leading to discrepancies in reporting.

193 The Code of Practice did not have commonly agreed definitions. Many of the stakeholders involved in mitigating disinformation use their own terminologies to encapsulate the problem, including information warfare, influence operations, hybrid warfare, coordinated inauthentic behavior, computational propaganda.

194 Illustrative examples of KPIs can be found in Chapter 5.2 for each proposed standard.

Accountability requires four key components: organizational and reporting transparency, meaningful independent oversight, adherence to fundamental rights, and remediation mechanisms.

and provide relevant information. In addition, the oversight board would also be able to evaluate the relevance of KPIs and suggest changes to them.

Second, it is through these transparency mechanisms that the oversight board is able to monitor the implementation and effectiveness of the standards to make sure they achieve the envisioned levels of transparency and hold the platforms accountable. Any lags or failures to meet the desired levels of transparency can be dealt with using the legislative base as a credible enforcement mechanism. Third, the regulator and multistakeholder body should include a strong protection of fundamental human rights since. Fourth, accountability requires access to remediation, meaning that if a piece of content is taken down, platforms should have a process in place through which users can challenge that decision. When platforms implement these standards into their policies, they should each offer a remediation process for users.

## 6.6 Institutionalizing coregulation: A short proposal for a DISINFO-ISAC

Harmful content like disinformation travels across platforms. Despite the networked nature of the actors, the content and its dissemination, threat information sharing on disinformation among social media platforms and with civil society and government merely occurs on an informal ad-hoc basis. With the aim to create a more coherent and standardized public-private alternative, a European Disinformation Information Sharing and Analysis Center (DISINFO-ISAC) is proposed. Its main goal is to facilitate threat intel sharing primarily among social media platforms from which European civil society and government institutions can also benefit.

### What? Public-private threat information-sharing and capacity building

Public-private Information Sharing and Analysis Centers (ISACs), sometime known as a Warning, Advice, and Reporting Point (WARP), have formed the backbone of national cybersecurity efforts for nearly 20 years. Today, national cybersecurity would be unthinkable without them (see Appendix IV for a thorough explanation on the role and typologies of ISACs). Building on this experience, we propose a Disinformation ISAC (DISINFO-ISAC) as a distinct instrument to facilitate information exchange primarily between social media platforms, but also with government agencies, and civil society with regular updates on potential threats, disinformation campaigns, delivery agents. It would also improve the capacity of smaller industry members.

### Who? Industry-led with civil society and government involvement

While different formats exist for ISACs, a bottom-up approach led by industry partners is most conducive for establishing a coalition of the willing that is guided by mutual trust and contributes to high-quality credible information sharing on a consistent basis. After all, the most mature and successful ISACs to this date, like those for the financial sector, have been organized in a bottom-up fashion, rather than through a top-down model.

1. **Social media platforms** are the primary driving force of an ISAC. At the beginning, the very large social media platforms will be required to do most of the labelling and sharing of information. After all, smaller platforms do not always have the capacity or resources to

Public-private Information Sharing and Analysis Centers (ISACs), sometime known as a Warning, Advice, and Reporting Point (WARP), have formed the backbone of national cybersecurity efforts for nearly 20 years.

monitor, track, cross-check and label disinformation content on their platforms. The ISAC would provide them with crucial threat intelligence and have a capacity-building function for all members, thereby increasing the overall expertise and security of the community so the least intrusive measures can be taken in respect to civil and human rights.

2. **Civil society and news agencies** would primarily function as consumers of information, at least in the beginning. There is a possibility, however, for them to take up a more proactive role in providing crucial intelligence or analysis outside of the social media platforms, displaying tangible multistakeholder cooperation. Civil society organizations could also contribute to audit the process, as well as review or, if necessary, appeal decisions through the ombudsman, providing much needed transparency.
3. **Governments'** role may be in facilitating the ISAC through hosting or funding, and second, in creating a legal framework or mandate for the establishment of the ISAC and information sharing. Public administration, like the EEAS or national institutions, can also actively participate in the information sharing and analysis functions, but are likely to be consumers of information. Involvement of the EEAS or the Canadian Secretariat of the G7 Rapid Response Mechanism would offer real-time updates from social media platforms to these government-to-government mechanisms, as well as allow them to communicate their priorities and goals to industry partners, thereby contributing to much-needed public-private cooperation in this field.

### **How? A STIX-TAXII for disinformation and capacity building**

In terms of information exchange, members of the DISINFO-ISAC would primarily track and share information on both content and on delivery agents, where the benefits would not only accrue to the large platforms themselves, but also the many smaller fora (such as online newspaper comments section) that play an important role in the political ecosystem in many European Member States.

### ***Disinformation information exchange & remediation***

This basis function would depend on a formal process similar to the GIFCT's Hash-Sharing Initiative, where known hashes of terrorist videos and content are shared between industry members. Hashing, previously known as PhotoDNA, is a digital fingerprinting technology for tagging violative content that is then shared in an anonymized way with partners so they can quickly identify and decide to take appropriate measures against it. The only content the GIFCT database exchanges is linked to a UN-listed terrorist entity or when a Content Incident Protocol (CIP) is activated.<sup>195</sup> While hashing works best for video and picture content, it is theoretically possible to apply the process to text-only as well, albeit with reduced accuracy.<sup>196</sup>

<sup>195</sup> In case of a CIP, the GIFCT creates a separate label within the database for the perpetrator-related content to be shared. The CIP was activated after the 2019 Christchurch attack, recognizing the need to be able to share hashes of content in the low-prevalence high-risk scenario of a real-world crisis. It has only been used twice thus far, in Halle, Germany in October 2019, and in Glendale, Arizona in May 2020. While the UN List and the CIP have helped companies scale-up efforts to surface violating content on their own platforms and apply their terms of service accordingly, GIFCT recognizes that this is a limited subset of terrorist and violent extremist content.

<sup>196</sup> For much of the terrorist or online child abuse material, the individual piece of content (picture or video) is the problem. For many of the disinformation campaigns, however, we are dealing with a much wider narrative that goes beyond one piece of content. Hashes can be used to tackle individual pieces of disinformation, but they won't be able to target different content pushing a similar narrative, although it may inform algorithms that can.

Large platforms regularly track and label content – sometimes even just individual text passages – according to their veracity, effectively providing a “fake ratio”. The DISINFO-ISAC would provide a platform to encourage the large social media companies to exchange their classifications according to a mutually intelligible format. The technical example here is from cyber threat intelligence sharing, where standards such as STIX/TAXII and Snort (among others) allow technical information to be communicated irrespective of platform or language. Currently there are no yet accepted threat intelligence formats although a number of models are rumored to be explored.

### STIX/TAXII and Snort

STIX or Structured Threat Information eXpression is a standardized language for describing cyber threat information, designed to be shared via TAXII, short for Trusted Automated Exchange of Indicator Information. STIX is an XML-based language and serialization format that is structured in such a way that it can describe threat information, such as the motivations, abilities, capabilities and the response, that it is understandable to all humans (regardless of their language) and machine users.<sup>197</sup> TAXII is an application layer protocol that defines how cyber threat intelligence, like a STIX package, is shared. STIX/TAXII are widely adopted in ISACs and were specifically designed to support each other but are independent standards - the structures and serializations of STIX do not rely on any specific transport mechanism and TAXII can also be used to transport other data.

Snort is an open-source intrusion detection system (IDS) for detecting malicious activity on computer networks.<sup>198</sup> It is the high-tech equivalent of a fire alarm that has become the de-facto standard IDS.<sup>199</sup> Detection systems like Snort form the second security portal after the firewall and is what an anti-virus system does for files: it inspects the content of the traffic on networks and looks for possible malicious activity.<sup>200,201</sup>

Developing or helping to contribute to the development of a “STIX/TAXII for disinformation” would be a major contribution to European norm-setting. It would also allow additional parameters to be added or discarded depending on the use – for instance, some companies might add tags that are geographically relevant (for instance including illegal content described under the German NetzDG Act). It would also greatly assist the problem of confronting disinformation on smaller social media companies or news agencies, such as the comments sections in major newspapers that plays an important role in the wider disinformation ecosystem of many European countries like Germany, France, Italy and Austria. These news agencies would largely just be consumers of the threat information data rather

197 Panos Kampanakis, “Security Automation and Threat Information-Sharing Options,” *IEEE Security & Privacy* 12, no. 5 (2014): 42–51. DOI: 10.1109/MSP.2014.99.

198 Syed Ali Raza Shah and Biju Issac, ‘Performance Comparison of Intrusion Detection Systems and Application of Machine Learning to Snort System’, *Future Generation Computer Systems* 80 (1 March 2018): 157–70, <https://doi.org/10.1016/j.future.2017.10.016>.

199 Shah and Issac, “Performance Comparison of Intrusion Detection Systems and Application of Machine Learning to Snort System.”

200 Brian Caswell, Jay Beale and Andrew Baker, *Snort Intrusion Detection and Prevention Toolkit* (Elsevier Inc, 2006), <https://doi.org/10.1016/B978-1-59749-099-3.X5000-9>.

201 Dongyan Zhang and Shuo Wang, “Optimization of Traditional Snort Intrusion Detection System,” *IOP Conference Series: Materials Science and Engineering* 569, no. 4 (2019). <https://doi.org/10.1088/1757-899X/569/4/042041>.

than providers, however they could also provide crucial intelligence. It would also provide the ability for other information providers than social media companies to provide input into the system, like academic or civil society researchers, and offer an ability to audit and if necessary, appeal false classifications. This provides much needed transparency for what is often a black box process. An important part of the process would also be auditing remediation for falsely tagged content. While the DISINFO-ISAC should likely not be the primary labeler of disinformation, it can pass on complaints and concerns to the right body, and equally important follow up if there is a lack of response. It is even possible for the body to incorporate an ombudsman for disinformation, a key intermediary between major platforms and the average user, although this could quickly become a colossal task and should be approached with caution and a clear mandate.

The information provided through the DISINFO-ISAC would be for the social media platforms – big or small – to act on as they see fit. No user content would *automatically* be added or removed through this system, although it could support an automatic tagging of suspicious content for possible annotation (as false or similar) or outright removal. While most of the tagging (and ranking) will be done by the larger platforms at least in the initial stages, over time civil society organizations and academia could be encouraged to play a more proactive role as well. To a certain extent the DISINFO-ISAC could also do this directly, although only for the 'highest' labels (most likely to be false information) and with proper attribution. It could also consider using a similar threshold for the triggering the RAS alert system: a disinformation campaign that has 'translational significant impact'.

### ***Delivery agent tracking and information exchange***

It would be better to separate the information exchange of disinformation providers from the disinformation content. A label on foreign disinformation actors would therefore be added, specifically targeting foreign state-owned or government-controlled or affiliated propaganda. Ideally this would result in a database of disinformation producers, both for human accounts as well as botnets. It would require natural language processing and unified standards and language around government-controlled and government-aligned media. There currently is no equivalent to the UN list of terrorist groups in the disinformation context, so exceptional care would have to be taken in the wording of the various categories explaining government involvement to avoid false equivalence between outlets which are editorially independent from governments but receive funding from them, such as the BBC, and outlets closely aligned with government policies, such as Russia Today. If done well, it should inform users of the source of information without unduly harming legitimate journalistic outlets with government ties.<sup>202</sup> Integrating this with 'STIX/TAXII for disinformation' would be possible – STIX for example already focuses on the identification of the attacker, and their tactics, techniques, and procedures (TTP).

### ***Capacity building***

Capacity building allows members to increase the overall expertise and security of the community as a whole and contribute to the level and quality of participation of smaller members. The experience with technical ISACs have shown that a major part of their taskset is supporting smaller members as well as educating external stakeholders, like in the media

<sup>202</sup> Michael McFaul, *Securing American Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond*, (Stanford: Stanford Cyber Policy Center, June 2019), <https://fsi.stanford.edu/publication/securing-american-elections-prescriptions-enhancing-integrity-and-independence-2020-us>.



and public policy, as to their role. As the case study on the GIFCT has shown, the development of automated content tagging tools at scale is very hard and resource intensive, requiring large datasets that are not available to smaller providers. One of the GIFCT's successes was that smaller companies were able to take less drastic solutions for identifying and removing terrorist propaganda. It also makes these smaller companies better equipped to resist governmental pressure to remove speech in that they can point to a tested procedure.

### Link to existing initiatives

The DISINFO-ISAC would facilitate threat information sharing similarly to the G7 Rapid Response Mechanism<sup>203</sup> and EU Rapid Alert System (RAS)<sup>204</sup>. In contrast to these inter-governmental mechanisms, the ISAC would be industry-led while allowing civil society and government involvement. The European Digital Media Observatory (EDMO), on the other hand, is a multistakeholder initiative but focuses on factchecking, research and analysis and media literacy, rather than on threat information sharing.

According to the European Democracy Action Plan, online platforms should cooperate with the RAS contact points, in particular during election periods, to provide relevant and timely information. In a 2021 evaluation, the EU Court of Auditors, however, concluded that “there is no protocol establishing cooperation between the RAS and the online platforms” and recommended that the EEAS proposes “to the online platforms and the Member States a framework for cooperation between the RAS and the online platform” to improve participation.<sup>205</sup> As the RAS and G7 Mechanism remain a government-to-government model, the DISINFO-ISAC can be one way to bridge this gap between social media platforms and governments.

The DISINFO-ISAC would not replace or eliminate current information exchanges but would be one of the first proposal of its kind to offer a more standardized and formal threat information sharing between platforms first, and with civil society and governments secondly in the context of disinformation. Firstly, as of now, there is no such formal channel to share signals across social media platforms. This does not mean that cooperation is not taking place, but current threat intel sharing occurs ad-hoc rather than structured, and does not involve outside stakeholders. While ad-hoc informal cooperation can be an effective approach, collaboration also needs to be seen by outsiders, rather than relying on the judgement of individual companies. Secondly, by involving those other government and civil society stakeholders (e.g. the EEAS and the Canadian Secretariat of the G7 Mechanism), the RAS and G7 Rapid Response

203 The Group of Seven (G7) announced the Rapid Response Mechanism in June 2018 to respond to efforts of foreign actors seeking to “undermine our democratic societies and institutions, our electoral processes, our sovereignty and our security.” The Mechanism is facilitated by Global Affairs Canada and allows for real time sharing of threat assessments, best practices and lessons learned, and works towards more coordinated action among G7 members. Government of Canada, *Charlevoix commitment on defending democracy from foreign threats*, (Government of Canada, 2019), [https://www.international.gc.ca/world-monde/international\\_relations-relations\\_internationales/g7/documents/2018-06-09-defending\\_democracy-defense\\_democratie.aspx?lang=eng](https://www.international.gc.ca/world-monde/international_relations-relations_internationales/g7/documents/2018-06-09-defending_democracy-defense_democratie.aspx?lang=eng)

204 Similarly to the G7 Mechanism, the EU decided to establish a Rapid Alert System (RAS) against disinformation in March 2019 as a result of its Democracy Action Plan (Pillar II, action 3). Facilitated by the EEAS, it functions as a network of national contact points and a web-based digital platform through which EU institutions and member states can “provide alerts on disinformation campaigns in real-time through a dedicated technological infrastructure” in order to “facilitate sharing of data and assessment, to enable common situational awareness, coordinated attribution and response and ensure time and resource efficiency”. In a recent evaluation of the European Democracy Action Plan and the RAS, the EU Court of Auditors found that the RAS currently does not go beyond information sharing and has not yet issued alerts or led to coordinated common actions or attribution. EC, *On the European democracy action plan* (Brussels: 2020)

205 European Court of Auditors, *Disinformation affecting the EU: tackled but not tamed* (Luxembourg: European Court of Auditors, 2021), 45. [https://www.eca.europa.eu/Lists/ECADocuments/SR21\\_09/SR\\_Disinformation\\_EN.pdf](https://www.eca.europa.eu/Lists/ECADocuments/SR21_09/SR_Disinformation_EN.pdf)

As the RAS and G7 Mechanism remain a government-to-government model, the DISINFO-ISAC can be one way to bridge this gap between social media platforms and governments.

Mechanism would receive real-time threat information updates from the social media platforms, while at the same time allowing them to communicate their priorities and goals to industry partners, thereby contributing to much-needed public-private communication and coordination in this field. Based on those threat intel feeds, government partners can then combine the ISAC's threat information with their own classified information and operationalize this into mitigation guidance or useful information that non-cleared partners can use. Taken together, it can form a crucial step towards a common crisis-management mechanism or "break-the-glass protocol" in the fight against disinformation.

## 6.7 Key Takeaways

A coregulation model is proposed to advance the standards defined in Chapter 5 from formulation to implementation. It retains industry leadership in setting the standards but goes one step further than the self-regulation model by introducing enforcement and noncompliance mechanisms, as it backs up the industry standards with a statutory layer and independent oversight.

The regulator would establish the high-level principles, while a multi-stakeholder body consisting of representatives from government, industry and civil society translates these principles into standards codified in an industry charter. The charter can build on the norms, standards and key performance indicators provided in this report and elsewhere. The social media platforms must then implement these standards and an independent oversight board appointed by the Commission monitors and reports on the performance of the platforms. Based on these reports, the regulator can consider penalties for noncompliance.

A Disinformation Information Sharing and Analysis Center (DISINFO-ISAC) is proposed, consisting of social media platforms, government and civil society, to facilitate information exchange, remediation for consumers and capacity building amongst industry participants. The DISINFO-ISAC would provide a platform through which platforms primarily track and share information on both disinformation content and delivery agents using a mutually intelligible format (similar to the STIX/TAXII and Snort standards used in cyber threat intelligence sharing). As a first step, harmonized and consistent definitions must be developed, taking the differences and idiosyncrasies across platforms into consideration. These definitions not only concern the term disinformation, but also common views on sources, thresholds for triggering a response and the response itself, as well as the level of preparedness.

A properly implemented DISINFO-ISAC could also illustrate that a coregulation model is tangible and accessible to the wider population. It thus contributes to a much-needed community of trust for threat information sharing and a risk assessment process to counter disinformation. It would allow for faster, more coordinated, transparent and universally comprehensible cooperation in a highly sensitive field that is visible to outside stakeholders without becoming overly transparent. That is, without the risk of threat information, such as intrusion methods, being shared publicly and thereby informing the adversary. At the same time, it would advance public-private cooperation through civil society and government involvement and inform existing government-to-government initiatives, such as the EU Rapid Alert System and G7 Rapid Response Mechanism. More work on this idea should be encouraged.

# 7 Conclusion and Recommendations

The information environment, much like cyberspace, is a complex space that is inhabited by a wide range of stakeholders from government, private sector and civil society. The regime complex that counters disinformation is still in its infancy and encompasses a wide range of actors that all play a role and have their own standards, norms and processes. Governments have drafted national laws and established their own task forces, while also proposing intergovernmental legislative and normative proposals. Companies at each level of the tech stack (from social media platforms, cloud services providers, content delivery networks, domain registrars, to Internet Service Providers), are dealing with moderation in the context of disinformation and inauthentic behavior. Finally, civil society stakeholders, including the technical community and academia, propose mitigation measures on the basis of research about the threat actors and methods. Given the unique and autonomous nature of the involved regimes that often work at odds, it is unlikely that there can be one unilateral legal or normative solution that works for all stakeholders. Instead, there is a need for more norm coherence, irrespective of whether they are mainly technical, legal or political, which leads to the first conclusion of this report.

**First, there can be no one legal or normative solution to counter disinformation that would work for the entire regime complex. Rather than convergence, there is a need for more norm coherence across technical, legal and political initiatives.** Disinformation campaigns are characterized by their cross-platform nature – it is a distributed phenomenon that leverages a network of assets across social media platforms, while making use of networking infrastructure and routing services, and multiple levels of the tech stack, from social media platforms to Internet Service Providers. In and on itself, disinformation is not strictly illegal according to international law, although parts of it can be illegal when it overlaps with certain illegal activities, such as foreign interference or defamation. In absence of explicit legal and normative prohibitions, this report explores two possible avenues for developing norms against disinformation. Both occur at the international Whole of System level, one focuses on a government-to-government “big N Norms”, much like the cyber norms in the United Nations, while the other focuses on non-state “small n norms” developed on the basis of social media platforms’ standards and best practices.

**Second, governments should proceed carefully with a “big N norm” against disinformation. It can be framed around covert election interference and linked to the nonintervention principle. This would prohibit concerted Russian disinformation campaigns and covert influence operations aimed at undermining democratic processes, while allowing overt support for democratic processes and voices. Such a norm proposal, however, comes with risks.**

When Western states publicly argue in favor of prohibiting the spread of disinformation to interfere in the internal affairs of states, they need to reject Russian total noninterference proposals that favor sovereignty over universal human rights and propose multilateral – rather than a multistakeholder – Internet governance mechanisms. Any new norm proposal runs the risk of being hijacked to support this view; ultimately risking a move towards the Russian concept of *information security* and an intergovernmental discussion on what content and information should be allowed online, opening the door for human rights abusers to argue for sovereign control of information and crack down on dissenting voices via censorship.

**Third, an industry charter against disinformation should be advanced. On the basis of the current moderation practices of the major social media platforms in dealing with disinformation, eight overarching standards, including their requirements, risks and key performance indicators, can be considered for an industry charter and as part of the forthcoming update of the EU Code of Practice on Disinformation** (see Table 7). Taking into consideration the significant risks of a government-to-government norm development against disinformation, the second avenue of a Whole of System approach is considered: “small n norm” development by the social media platforms in the form of an industry charter to which we propose eight standards and KPIs that aim to incentivize transparency and serve to evaluate the effectiveness of the standard and the performance of the platforms. Taking into consideration the cross-platform nature of disinformation campaigns, the proposed standards also urge for greater coherency of efforts across platforms and for commonly agreed definitions concerning disinformation.

*Table 7 Eight proposed industry standards to be included in an industry charter.*

Standard	Description
<b>Community Guidelines</b>	Community guidelines and Terms of Service should clearly outline social media platforms' policy on disinformation, including definitions of what constitutes a violation and the corresponding consequences.
<b>Bot Takedowns</b>	Social media platforms should remove malicious bot accounts, botnets, or coordinated inauthentic behavior to ensure that only organic human activity is reflected in various measures of popularity, authority, and influence on social media. They are also encouraged to consider preventative measures that can include authenticity verification measures to help prevent these accounts from being created.
<b>Factchecking</b>	Social media platforms should have a factchecking process in place, preferably supported by an accredited third party that acts according to independently-established industry standards when labeling disinformation.
<b>Labelling</b>	Social media platforms should create standardized guidelines for a transparent, coherent, and multilevel labeling system dealing with (1) identified disinformation content (potentially with a ranking); (2) sponsored content (including political advertising); and (3) disinformation actors (including machine and human).
<b>Political Advertising</b>	Social media platforms should take a number of steps to clearly label sponsored content (including political advertising), including requiring verification from the sponsor and having a minimum data reporting requirement on their ad revenue streams. Along these lines, platforms should especially increase their support of current ad repositories to aid researchers. Second, platforms need to increase their oversight over political advertising, as well as limit the targeting capabilities for political advertising.
<b>Verified Information Features</b>	Social media platforms are encouraged to actively use verified information features, such as links to and pages of factchecked information that debunk disinformation during concerted campaigns or focal points, such as elections or the COVID-19 pandemic. Platforms should also be encouraged to apply such features to other societal issues, such as climate science denial, based on independent disinformation threat assessments. Finally, in exceptional circumstances, a platform should consider using its advertising algorithms to target victims of disinformation with verified information to actively debunk falsehoods.
<b>Algorithms and automated content moderation</b>	Social media platforms cannot rely on artificial intelligence alone for their moderation but must employ human moderators that are familiar with the local context and language, as well as establish efficient appeal procedures. Platforms need to be transparent about how algorithms work to both suggest and promote content, as well as how they are used in content moderation.
<b>Community Reporting and Remediation</b>	Social media platforms should have a dedicated community reporting mechanism for disinformation, take measures to timely mitigate reported disinformation and be transparent about their process. Similarly, platforms should guarantee that individuals have the ability to appeal a decision by the platform.

**Fourth, without more transparency from the platforms on their implementation and enforcement of the standards, there can be no meaningful oversight, accountability or insight into the effectiveness of each standard.** What you measure and report on creates incentives. The key performance indicators proposed in this report aim to go beyond measuring absolute numbers (e.g. the number of takedowns), and include measures that assess the quality of a platform's moderation process. This would help monitor implementation and provide evidence towards more comprehensive assessments of the standards' impact.

These KPIs should also be (re)evaluated periodically to ensure their compliance with the respective criteria. At the same time, there should be an understanding of the notion that, especially in the case of foreign influence operations, public transparency standards can actually be counterproductive when it comes to deterring malicious actors, as they could adjust their *modus operandi* in such a way to circumvent content moderation triggers and effectively game the system. This recommendation is grounded in an evaluation of various regulatory models (see Annex III), most notably the shortcomings of the self-regulation approach adopted by the EU Code of Practice on Disinformation in effecting meaningful transparency, accountability and civil-society participation, as well as the collective black box process of the centralized industry-cooperation model taken by the GIFCT (see Annex II).

**Fifth, resources can be directed towards an industry and expert wide survey evaluating the feasibility and effectiveness of the eight proposed standards and the KPIs.** While action can be reasonably achieved, the important question is whether such standards can be feasibly attained. A survey can evaluate the feasibility of implementing the proposed standards across different metrics, such as costs, incentives, current level of support and transparency.

**Sixth, the goal is to strike the right balance between the ideal end-goal of the standards and the feasibility of their wide implementation by the social media platforms.** Finding the most appropriate language to express a standard or norm can be challenging. If they are too precise, it may be hard to achieve consensus, relevancy for all platforms, and avoid gaps in coverage. If they are too vague, they do not provide concrete guidance or expectations. In addition, they cannot be static. Community guidelines of social media companies, for example, change very quickly, mostly to respond to a continually changing technology and threat landscape. Actors should be prepared to augment or adapt existing norms and standards and develop new ones as technologies and our understanding of their implications change. Standards should therefore strike the right balance between their ideal end-state and the likelihood of them being widely implemented. After all, for a standard to be effective, it must be adopted and implemented, and platforms need to be held accountable.

**Seventh, a European coregulation model should be developed to advance the industry charter from formulation to implementation. It retains industry leadership in setting the standards but goes one step further than the current self-regulation model in introducing a statutory layer and independent oversight to enforce the standards and create noncompliance mechanisms.** The regulator (e.g. European Commission) would establish the high-level principles, while a multi-stakeholder body consisting of representatives from government, industry and civil society translates these principles into standards codified in an industry charter. The charter can build on the norms, standards and key performance indicators provided in this report and elsewhere. The social media platforms must then implement these standards and an independent oversight board appointed by the Commission monitors and reports on the performance of the platforms. Based on these reports, the regulator can consider penalties for noncompliance. Building in such third-party oversight and accountability is crucial to avoid abuse by and lack of responsibility from social media companies. Importantly, oversight mechanisms need to be strong enough so red flags can be raised on the basis of sufficient information.

**Eighth, a Disinformation Information Sharing and Analysis Center (DISINFO-ISAC) can be established to contribute towards a much-needed community of trust for faster, more coordinated, and transparent threat information sharing across social media platforms, and with civil society and government.** The DISINFO-ISAC would not replace or eliminate



current information exchanges but would be one of the first proposal of its kind to offer a more standardized and formal threat information sharing on disinformation. As of now, such threat information sharing among industry partners occurs on an ad-hoc basis. The DISINFO-ISAC would be an industry-led initiative but through its involvement of civil society and government stakeholders, it does not only show it can be a forum for effective collaboration, but that such collaboration is seen, rather than relying on the judgement of individual companies. It would thereby improve public-private cooperation and trust, as well as inform related government-to-government initiatives, such as the EU Rapid Alert System (RAS) and G7 Rapid Response Mechanism. By involving those other stakeholders, the RAS and G7 Mechanism would receive real-time threat information updates from the social media platforms, while at the same time allowing them to communicate their priorities and goals to industry partners. Based on those threat intel feeds, government partners can then combine the ISAC's threat information with their own classified information and operationalize this into mitigation guidance or useful information that non-cleared partners can use. Taken together, it can form a crucial step towards a common crisis-management mechanism or "break-the-glass protocol" in the fight against disinformation.

**Ninth, developing a STIX/TAXII for disinformation standards would be a major contribution to threat information sharing irrespective of platform or language.** ISAC members would primarily track, share and label information on both disinformation content and delivery agents according to their veracity, effectively providing a 'fake ratio' and exchange their classifications according to a mutually intelligible format. The technical example here is from cyber threat intelligence sharing, where using standards such as STIX/TAXII and Snort (among others) allows threat intel to be communicated in a way that it is understandable to all humans (regardless of their language) and machine users. It would provide crucial threat intelligence to smaller social media companies and news agencies to act upon, while limiting the risk of it being shared publicly and thereby informing the adversary to change course or becoming weaponized by other malicious actors.

**Tenth, the DISINFO-ISAC facilitates capacity building among industry partners and possible remediation.** The experience with cybersecurity ISACs has shown that a major part of their taskset is supporting smaller members as well as educating external stakeholders. As a result, smaller companies would be able to take less drastic moderation solutions. It also makes them better equipped to resist governmental pressure to remove speech in that they can point to a tested procedure. Most of the tagging (and ranking) will be done by the larger platforms at least in the initial stages. Over time, civil society organizations and academia could be encouraged to give input into the system, and provide an ability to audit and, if necessary, appeal false classifications. Such a remediation process for falsely tagged content can be done by the ISAC or through an ombudsman, which passes on complaints to the responsible body and follows up if there is a lack of response.

These recommendations come at a time when the European self-regulatory approach towards social media companies' responsibility is slowly shifting towards coregulation. The proposed standards and key performance indicators for social media platforms, as well as the recommendations for accountability and multistakeholder engagement in coregulation are therefore timely. Most notably the DISINFO-ISAC would a major contribution to European norm-setting. We therefore call on the European Commission to consider these and the proposals made elsewhere to strengthen the EU Democracy Action Plan, improve the responsibility of social media platforms in countering disinformation, and strengthen their cooperation with other stakeholders.

# 8 Annex I: industry best practices for countering disinformation

This Annex offers a detailed overview of the current best practices of the major social media platforms (Twitter, Facebook, Google, Microsoft, TikTok) on content and conduct moderation and other counter-disinformation activities, including labelling, community or voluntary reporting, third-party factcheckers, oversight board, community guidelines, algorithmic and automated moderation, and verified information features. Each table describes the current measures taken by each company as reported by the platforms themselves and is followed by conclusions that inform the standards proposed in Chapter 5.

The measures described in this Annex are updated regularly by the social media platforms, so it should be noted that the collected data was collected up to April 2021. Unfortunately, most platforms do not maintain a centralized location with all policy documents on disinformation. Instead, they often publish using different formats (Terms of Service, Community Guidelines, web publications, policy documents, strategies, social media posts, or similar) without offering any guidance as to how documents relate to each other.

## 8.1 Labelling

### Twitter

Twitter has enforced labelling policy most aggressively compared to other platforms. Its first overt attempts at labelling started in June 2019, when the company announced it would start adding warning labels to tweets from political figures. These labels would allow the tweets to remain on the website, and were determined by Twitter employees on a seemingly case-by-case basis.<sup>206</sup> In February 2020, Twitter announced its policy for removing or labelling any tweets which “deceptively share synthetic or manipulated media that are likely to cause harm.”<sup>207</sup> Herein, labelling is used to “help people understand the media’s authenticity and to provide additional context.”<sup>208</sup> These tools were especially prominent during the 2020 American election cycle.<sup>209</sup> In May 2020, in light of the Covid-19 pandemic, this policy was expanded to also address, label, and/or remove content going against authoritative sources in terms of Covid-19 guidance.<sup>210</sup> These innovations have continued: Starting in 2021, Twitter announced that it “may label or place a warning on Tweets that advance unsubstantiated rumors, disputed claims, as well as incomplete or out-of-context information about vaccines.”<sup>211</sup>

Twitter currently applies labels for two categories, and is likely to introduce new ones:

1. “misleading information - statements or assertions that have been confirmed to be false or misleading by subject-matter experts, such as public health authorities”
2. “disputed claims – statements or assertions in which the accuracy, truthfulness, or credibility of the claim is contested or unknown”<sup>212</sup>

Whereas the disputed label still allows for normal interaction, the misleading label prompts a message that includes credible information before someone is able to retweet or further amplify the post, and by forcing retweets to go through the “quote tweet” user interface instead. This change aims to slow people down from quickly retweeting posts without adding their own commentary. Additional restrictions were also added specifically for accounts owned by US political figures with more than 100,000 followers: now, if one of their tweets gets flagged with a ‘misleading information’ label, people must tap through a warning screen in order to see that tweet. These changes were introduced in October 2020.<sup>213</sup>

206 Kate Conger, “Twitter to Label Abusive Tweets From Political Leaders,” *The New York Times*, June 27, 2019, <https://www.nytimes.com/2019/06/27/technology/twitter-politicians-labels-abuse.html>.

207 Yoel Roth and Ashita Achuthan, “Building rules in public: Our approach to synthetic & manipulated media,” *Twitter*, February 4, 2020, [https://blog.twitter.com/en\\_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html](https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html); Twitter, “Synthetic and manipulated media policy,” *Twitter Help Center*, accessed on May 11, 2021, <https://help.twitter.com/en/rules-and-policies/manipulated-media>; @Delbius. “Help us shape our approach to synthetic and manipulated media.” *Twitter*, November 11, 2019. [https://blog.twitter.com/en\\_us/topics/company/2019/synthetic\\_manipulated\\_media\\_policy\\_feedback.html](https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html).

208 Twitter, “Synthetic and manipulated media policy.”

209 Kate Conger, “Twitter says it labeled 0.2% of all election-related tweets as disputed,” *The New York Times*, November 12, 2020, <https://www.nytimes.com/2020/11/12/technology/twitter-says-it-labeled-0-2-of-all-election-related-tweets-as-disputed.html>; Kate Conger, “Twitter Had Been Drawing a Line for Months When Trump Crossed It,” *The New York Times*, May 30, 2020, <https://www.nytimes.com/2020/05/30/technology/twitter-trump-dorsey.html>.

210 Yoel Roth and Nick Pickles, “Updating our approach to misleading information,” *Twitter*, May 11, 2020, [https://blog.twitter.com/en\\_us/topics/product/2020/updating-our-approach-to-misleading-information.html](https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html).

211 Ibid.

212 Ibid.

213 Vijaya Gadde and Kayvon Beykpour, “Additional steps we’re taking ahead of the 2020 US Election,” *Twitter*, October 9, 2020, [https://blog.twitter.com/en\\_us/topics/company/2020/2020-election-changes.html](https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html).

## Facebook

Facebook began experimenting with labelling in the wake of the 2016 US elections.<sup>214</sup> Announced in December 2016 for American users, certain posts (if reported enough) would be factchecked by a third-party factchecker. They could then attach a red label to the content, limited to labelling the content as 'disputed'. This label then contained a link to 'additional information' explaining the decision. In 2017, the 'disputed' flagging was removed in favor of a direct link to an authoritative news source, with Facebook citing the 'Disputed' label as actually reinforcing beliefs.<sup>215</sup>

Throughout 2020, Facebook made a number of revisions and changes to its labelling program. First announced in October 2019<sup>216</sup> and implemented in June 2020<sup>217</sup>, Facebook and Instagram<sup>218</sup> now label all pages and ads produced by 'state controlled media.' It is unclear whether this label has any impacts other than informational.<sup>219</sup> Next, in response to the Covid-19 pandemic, Facebook announced in April 2020 that "Claims about COVID-19 or vaccines that do not violate these policies will still be eligible for review by our third-party fact-checkers, and if they are rated false, they will be labeled and demoted."<sup>220</sup> Facebook expanded this in March 2021 to add informational labels to all posts about vaccines.<sup>221</sup> Finally, Facebook also updated its misinformation labels in 2020, mainly in responses to the 2020 American elections. Now, these labels specify the type of falsehood in the content, as determined by the factchecking program. Labels include "Altered", "Missing Context", "Partly False", and "False".<sup>222</sup> These labels have different magnitudes of actions:

Label	Action
<b>Missing Context</b>	Facebook will introduce a "lighter-weight warning label" and "...focus on surfacing more information from our fact-checking partners." <sup>223</sup>
<b>Partly False</b>	Facebook will introduce a "lighter-weight warning label" and "...reduce the distribution of this content, but to a lesser degree than "False" or "Altered." <sup>224</sup>
<b>Altered</b>	"[Facebook will]...dramatically reduce the distribution of these posts, and apply our strongest warning labels." <sup>225</sup>
<b>False</b>	An internal Facebook study, however, reported that labelling was not very effective, most likely because there are no limitations linked to the label. <sup>226</sup> This is somewhat evident in Facebook's reports to the European Commission surrounding its response to the Covid-19 disinformation crisis <sup>227</sup> , where some of them do not mention the term 'labelling' at all, preferring to focus on quasi-labelling initiatives such as informative pop-ups on their news feed <sup>228</sup> and misinformation warning screens. <sup>229</sup>

214 Adam Mosseri, "Addressing Hoaxes and Fake News," Facebook, December 15, 2016, <https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/>; Alex Heath, "Facebook is going to use Snopes and other fact-checkers to combat and bury 'fake news'," *Business Insider*, December 15, 2016, <https://www.businessinsider.nl/facebook-will-fact-check-label-fake-news-in-news-feed-2016-12?international=true&r=US>.

215 Catherine Shu, "Facebook will ditch Disputed Flags on fake news and display links to trustworthy articles instead," *Techcrunch*, December 21, 2017, <https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/>.

216 Guy, Rosen, Katie Harbath, Nathaniel Gleicher, and Rob Leathern. "Helping to Protect the 2020 US Elections," Facebook, October 21, 2019, <https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/>.

217 Adi Robertson, "Facebook starts labeling 'state-controlled media' pages," *The Verge*, June 4, 2020, <https://www.theverge.com/2020/6/4/21280542/facebook-state-controlled-media-account-post-label-election-interference-ads-rt>.

218 Ibid.

219 Nathaniel Gleicher, "Labeling State-Controlled Media On Facebook," Facebook, June 4, 2020, <https://about.fb.com/news/2020/06/labeling-state-controlled-media/>.

220 Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," Facebook, April 16, 2020, <https://about.fb.com/news/2020/04/covid-19-misinfo-update/>.

221 Elizabeth Culliford, "Facebook to label all posts about COVID-19 vaccines," *Reuters*, March 15, 2021, <https://www.reuters.com/article/us-health-coronavirus-facebook-idUSKBN2B70NJ>.

222 Keren Goldshlager and Aaron Berman, "New Ratings for Fact-Checking Partners," *Facebook Journalism Project*, August 11, 2020, <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/new-ratings>.

223 Ibid.

224 Ibid.

225 Ibid.

226 Jen Patja Howell, "The Lawfare Podcast: Alex Stamos on Fighting Election Disinformation in Real Time," *Lawfare*, August 20, 2020, <https://www.lawfareblog.com/lawfare-podcast-alex-stamos-fighting-election-disinformation-real-time>; Facebook, *Facebook response to the European Commission Communication on Covid-19 Disinformation: Report for December 2020* (Facebook: 2021).

227 Facebook, *Facebook response to the European Commission Communication on Covid-19 Disinformation: Report for December 2020*.

228 Ibid., 1.

229 Ibid., 4.

## Google

First developed in 2016 for the American election, Google partners with websites like Snopes and PolitiFact to apply labels to news results in Google Search.<sup>230</sup> This was expanded in 2017 to all languages.<sup>231</sup> Announced in June 2020, Google will start adding labels to certain factchecked images in Google Search.<sup>232</sup> These labels are run using the Schema.org ClaimReview markup, which allows publishers to make public fact check statements. As such, Google emphasizes that “These fact checks are not Google’s and are presented so people can make more informed judgements.” Despite this, Google does propose standards for factcheckers.<sup>233</sup> Moreover, Google uses algorithms to choose whether or not to implement factchecker’s submitted labels, and therefore does not guarantee that factchecks or labels will show up even if the factcheckers appears to meet all the guidelines.<sup>234</sup>

First started in 2009, Google has long used labels including “In-Depth,” “Opinion,” “Blog,” and “Local Source” on news articles and search results.<sup>235</sup> In June 2017, it also implemented a “Factcheck” label which indicates the article has been factchecked.<sup>236</sup> Google also applies labels to political advertisers, with them reportedly labelling 444,000 ads for the 2020 European elections.<sup>237</sup>

Google’s video sharing platform, YouTube, also includes some efforts at labelling. For instance, in 2018, YouTube became the first major platform to label content from state-funded broadcasters,<sup>238</sup> stating that “The notice will appear below the video, but above the video’s title, and include a link to Wikipedia so viewers can learn more about the news broadcaster.”<sup>239</sup> However, the effectiveness of these labels have been questioned by experts: some studies show that they only work if explicitly and clearly placed.<sup>240</sup> Identifying disinformation in lengthy videos is, after all, markedly more challenging than for written media. Labelling is therefore especially difficult on a video-sharing platform: not only is disinformation more difficult to catch; but there are also more issues with where and when to display labels. For instance, YouTube does not highlight the particular section of the video where disinformation is located, only the video in general. But, having for instance a pop-up whenever disputed content is said, while effective, would also be extremely invasive.

230 April Glaser, “Google is rolling out a fact-check feature in its search and news results,” *Recode*, April 8, 2017, <https://www.vox.com/2017/4/8/15229878/google-fact-check-fake-news-search-news-results>.

231 Justin Kosslyn and Cong Yu, “Fact Check now available in Google Search and News around the world,” *Facebook*, April 7, 2017, <https://blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/>.

232 Harris Cohen, “Bringing fact check information to Google Images,” *Facebook*, June 22, 2020, <https://www.blog.google/products/search/bringing-fact-check-information-google-images>; Rachel Lerman, “Seeing isn’t always believing: Google starts fact-checking images,” *The Washington Post*, June 23, 2020, <https://www.washingtonpost.com/technology/2020/06/22/google-fact-check-images/>.

233 Kosslyn and Yu, “Fact Check now available in Google Search and News around the world.”

234 Google, “Fact Check,” Last updated: March 18, 2021, <https://developers.google.com/search/docs/data-types/factcheck>.

235 Rawan Hakeem, “Highlighting the diversity of content in Google News,” *Google*, September 17, 2009, <https://news.googleblog.com/2009/09/highlighting-diversity-of-content-in.html>.

236 Jaikumar Vijayan, “Google Introduces Fact Check Label on News Stories, Search Results,” *eWeek*, April 7, 2017, <https://www.eweek.com/cloud/google-introduces-fact-check-label-on-news-stories-search-results/>; Richard Gingras, “Labeling fact-check articles in Google News,” *Google News Initiative*, October 13, 2016, <https://blog.google/outreach-initiatives/google-news-initiative/labeling-fact-check-articles-google-news/>.

237 EC, *Assessment of the Code of Practice on Disinformation*, 5.

238 Thuy Ong, “YouTube will start labeling videos from state-funded broadcasters,” *The Verge*, February 2, 2018, <https://www.theverge.com/2018/2/2/16964190/youtube-state-funded-broadcasters>; Nassetta and Gross, “State media warning labels can counteract the effects of foreign misinformation.”

239 Geoff Samek, “Greater transparency for users around news broadcasters,” *You Tube*, February 2, 2018, <https://blog.youtube/news-and-events/greater-transparency-for-users-around>.

240 Nassetta and Gross, “State media warning labels can counteract the effects of foreign misinformation.”

## Microsoft

As announced in 2017, “Bing is adding a new UX [User Experience] element to the search results, called the “Fact Check” label, to help users find factchecking information on news, and with major stories and webpages within the Bing search results. Bing may apply this label to any page that has schema.org ClaimReview markup included on the page.”<sup>241</sup>

Microsoft also partners with organizations like NewsGuard, which offers browser plugins which add ‘nutrition labels’ to news websites encountered.<sup>242</sup> This plugin attaches a label to webpages which, much like nutrition labels operate on food, shows the consumer the credibility of a website or news source (refer to Figure 6). NewsGuard also includes links to more detailed write-ups on their criteria for websites and sources. In May 2020, Microsoft announced that all users of their Edge browser are able to install this software for free.<sup>243</sup>





RATING CATEGORIES	
	<b>Green:</b> A website is rated green if it generally adheres to basic standards of credibility and transparency. (If the site adheres to all nine of our criteria, we note that in the rating. If it has significant exceptions among the criteria, we note that too.)
	<b>Red:</b> A website is rated red if it generally fails to meet basic standards of credibility and transparency. (We note whether the site failed several of our nine criteria, or whether it severely violates journalistic standards by failing an especially significant number of criteria.)
	<b>Satire:</b> A humor or satire site receives a satire rating, indicating that it is not a real news website. We do not rate these sites according to the nine journalistic criteria, but we do provide a description of each site including, if possible, who is behind it.
	<b>Platform:</b> A site receives a platform rating if it primarily hosts user-generated content that it does not vet. Information from platform sites may or may not be reliable. We do not rate these sites according to the nine journalistic criteria, but we do provide a description of each site and its practices.

Figure 6: NewsGuard Rating Categories; from <https://www.newsguardtech.com/ratings/rating-process-criteria/>

241 Schema.org is a project started by Google, Microsoft, Yahoo, and Yandex which uses technologies like AI and Machine learning to attach source labels to websites and webpages to create, maintain, and promote schemas for structured data. One of their projects is Claim Review, which ‘tags’ information from certain sources as accurate, thus allowing search engines to promote certain sources over others on terms of their accurateness. Microsoft Bing Blogs, “Bing adds Fact Check label in SERP to support the ClaimReview markup,” *Microsoft*, September 14, 2017, <https://blogs.bing.com/Webmaster-Blog/September-2017/Bing-adds-Fact-Check-label-in-SERP-to-support-the-ClaimReview-markup#content>; RAND, “Schema.org Claim Review,” accessed May 11, 2021, <https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search/items/schemaorg-claim-review.html>; Schema.org, “ClaimReview,” last updated March 8, 2021, <https://schema.org/ClaimReview>.

242 Tom Burt, “New Steps to Combat Disinformation,” *Microsoft*, September 1, 2020, <https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/>.

243 NewsGuard, “Microsoft Expands NewsGuard Adoption,” May 14, 2020, <https://www.newsguardtech.com/press/microsoft-expands-newsguard-adoption/>.



## TikTok

In February 2021, TikTok announced that it was introducing warning labels for videos which potentially contain misinformation, which warn the user the TikTok has not been verified while viewing and before sharing.<sup>244</sup> Similar labels were first implemented on videos relating to Covid-19 vaccines in December 2020, which directed users to authoritative information.<sup>245</sup> As described by TikTok in January 2021, “We have also been rolling out a new vaccine tag to detect and tag all videos with words and hashtags related to the COVID-19 vaccine. We attach a banner to these videos with the message ‘Learn more about COVID-19 vaccines’. This then redirects the user to verifiable, authoritative sources of information.”<sup>246</sup>

### Labelling is a content moderation tool that is already implemented by the major platforms.

It has become one of the key responses to disinformation by platforms, and studies have shown that labelling can be an effective tool in decreasing the spread of disinformation.<sup>247</sup> While many platforms are still new to labelling (for instance, TikTok unveiled its labelling system in February 2021), the fact that all platforms have such measures should be seen as an endorsement of labelling as a feasible and attainable counter-disinformation standard.

**Coherence between platforms on labelling is scarce.** Platforms use their own labels and do not often collaborate with their peers when creating labels. There are also technical issues associated with labels. As some observers have noted, when content is shared across platforms (or occasionally even on the same platform), the labels are not.<sup>248</sup> Collaboration would greatly improve both the effectiveness and coherence of labels for consumers in several areas.

Having uniformity in the labels across platforms would help users more rapidly recognize the respective labels over time and allow them to become familiar with them more quickly, making labels more easier to process and also appear more trustworthy and even legitimate.<sup>249</sup> It would also facilitate better platform cooperation in dealing with disinformation that moves across different platforms. A common label would also help platforms avoid pitfalls such as attracting unnecessary attention to the mis-/disinformation, inconsistent labelling, not repeating falsehoods, and using non-confrontational language.<sup>250</sup> This obviously requires the platforms to first of all agree on what type of content should be addressed with labels,

244 No Author, “TikTok Introduces Warning Label To Combat Fake News,” *Entrepreneur Europe*, February 4, 2021, <https://www.entrepreneur.com/article/364767>; Gina Hernández, “TikTok añade nuevas indicaciones que ayudan a reconsiderar antes de compartir,” *TikTok*, accessed May 11, 2021, <https://tiktok.prezly.com/tiktok-anade-nuevas-indicaciones-que-ayudan-a-reconsiderar-antes-de-compartir/>; Whitney Kimball, “TikTok Is Adding a Potential Misinformation Warning Label to Save Us From Ourselves,” *Gizmodo*, February 3, 2021, <https://gizmodo.com/tiktok-is-adding-a-potential-misinformation-warning-lab-1846189941>.

245 Martyn Landi, “TikTok adds new vaccine misinformation labels and strengthens community rules,” *Breaking-news.ie*, December 15, 2020, <https://www.breakingnews.ie/business/tiktok-adds-new-vaccine-misinformation-labels-and-strengthens-community-rules-1051447.html>.

246 TikTok, *December 2020 Report EU Code of Practice on Disinformation / COVID-19* (TikTok: 2021), 5.

247 Peter Dizikes, “The catch to putting warning labels on fake news,” *MIT News*, March 2, 2020, <https://news.mit.edu/2020/warning-labels-fake-news-trustworthy-0303>; Pennycock et al, “The Implied Truth Effect.”

248 The ERGA noted that “in Facebook, if a political ad is shared by a user, the “Paid for by” disclaimer vanishes because the content is seen as organic by Facebook. This latter finding is very interesting, as it shows an important limitation to the effectiveness of the system.” European Regulators Group for Audiovisual Media Services, *ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice* (ERGA: 2020), 19. A similar issue also exists with YouTube: much of their videos, when shared on external platforms like Facebook, also do not get labelled properly. Aleksi Knuutila, Aliaksandr Herasimenka, Hubert Au, Jonathan Bright, Rasmus Nielsen, and Philip N. Howard, “COVID-related Misinformation on YouTube: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies,” *Oxford Internet Institute* (2020), <https://demotech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/09/YouTube-misinfo-memo.pdf>.

249 Saltz et al, “It matters how platforms label manipulated media. Here are 12 principles designers should follow.”

250 Ibid.

something which changes quite quickly for each platform. For instance, YouTube, Facebook, and Twitter apply labels to state-funded media,<sup>251</sup> while other platforms like Microsoft ban this type of content altogether. This, like the labelling design mentioned before, can be confusing to the user. Ideally, each platform could agree on what content should be labelled, and that should be coherent across platforms.

While labelling is an area of focus for the Codes of Practice on Disinformation, coherency between labels is not emphasized. Governments are therefore encouraged to be more proactive in pushing for platform collaboration in labelling practices. This should include establishing (a) common labels and (b) common standards for what content should be labelled.

**Labels should be linked to additional actions that affect the interaction with content.** Much like Twitter's labelling policy, platforms can use additional measures for the top-tiered labels. This can include a link to factchecked information and an advanced warning before or limitations on sharing to slow users down from quickly sharing disinformation.

**Platforms need to share more data to ensure transparency over the effectiveness of measures, including labelling.** It remains very difficult to get a practical insight into how helpful labelling mechanisms actually are to prevent disinformation. Studies can show that labelling works at reducing the spread of disinformation,<sup>252</sup> but without actual statistics from platforms themselves it becomes much more difficult to comment on their real-world impact. The European Commission noted the need for transparency on platform measures as a key takeaway from their review of the EU Code of Practice on Disinformation, stating that for tools like labels, "no data has been made available to demonstrate the extent to which such tools are effective to increase user engagement with trustworthy information sources, enhance critical thinking, and promote civic behavior online."<sup>253</sup> Simply put, without the proper transparency from platforms on how effective labels are, we cannot know how good of a solution they are for disinformation. Governments should especially focus on obtaining this information to better inform whether labelling should deserve a large focus in future anti-disinformation campaigns and initiatives.

**There are risks with labelling, most notably the "implied truth effect", that can be mitigated by additional 'true' or 'unchecked' labels.** Based on the "backfire effect",<sup>254</sup> the "implied truth effect" states that since labelling can never tag 100% of the platform's content, *false content that fails to get tagged can be viewed as truthful and thus are seen as more accurate by consumers than if there were no labels at all.*<sup>255</sup> Overall, this risk does not undermine the credibility of labeling as a whole, especially since researchers have examined feasible mitigation methods, such as adding 'true' content labels<sup>256</sup> or labelling all content as 'unchecked' by default.<sup>257</sup>

<sup>251</sup> Nassetta and Gross, "State media warning labels can counteract the effects of foreign misinformation."

<sup>252</sup> Dizikes, "The catch to putting warning labels on fake news,"; Pennycock et al, "The Implied Truth Effect."

<sup>253</sup> EC, *Assessment of the Code of Practice on Disinformation*, 10.

<sup>254</sup> The "backfire effect" is a phenomena when "individuals who receive unwelcome information ... may come to support their original opinion even more strongly." Nyhan and Riefler, "When Corrections Fail," 9.

<sup>255</sup> Pennycock et al, "The Implied Truth Effect."

<sup>256</sup> Ibid.

<sup>257</sup> Saltz et al, "It matters how platforms label manipulated media. Here are 12 principles designers should follow."

**Consumers can view labels as a form of censorship of their freedom of speech. Platforms should therefore try to make sure that labels are employed transparently, ideally through third-party factcheckers, and reasonably, through carefully balancing the intrusiveness of measures.** A 2020 working paper by Emily Saltz, Claire Leibowicz, and Clare Wardle studied the perspectives of consumers on labels. They found that “many [interviewed participants] viewed labeling as judgmental, paternalistic, and against the platform ethos.”<sup>258</sup> They identified three types of issues individuals tended to have with labelling: (1) False positives: citizens tend to not like it when content they do not view as offensive get labelled, (2) Overlays as censorship: some viewed labels as a form of censorship, and (3) The mythical ‘unbiased’ label: many expressed a desire for a ‘neutral’ authority to be in charge of labelling, and not platforms who they perceive as having an agenda.<sup>259</sup>

This issue is difficult to solve: after all, labelling is meant to guard against those using speech too liberally. Moreover, many particularly ideologically motivated individuals will balk at any mention of a label, no matter how non-intrusive. Yet, there are some measures which have been explored to mitigate these concerns. In terms of intrusiveness, labels could be carefully designed to exist in between too explicit and too subtle.<sup>260</sup> In terms of the bias, the transparency behind who places those labels can be increased, with a default preference towards third-party factcheckers, thereby encouraging the role of civil society and increasing its cooperation with platforms. Some studies have explored how to achieve a functional balance between these competing interests: Clare Wardle and others proposed 12 cogent principles on this following an analysis of best practices, which can be seen in Figure (7).



Figure 7: Twelve Design Principles for Labelling Media, Partnership on AI, 2020.

<sup>258</sup> Emily Saltz, Claire Leibowicz, and Claire Wardle. “Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions,” (December 2020): 8.

<sup>259</sup> Ibid., 9.

<sup>260</sup> Saltz et al, “It matters how platforms label manipulated media. Here are 12 principles designers should follow.”

## 8.2 Community or Voluntary Reporting

### Twitter

Twitter has developed a feature allowing certain users to report electoral misinformation or manipulation. As they wrote in 2019, a user can select ‘report Tweet’ from a dropdown menu on any tweet to do so. They can then submit a report outlining why they believe the tweet is misinformation.<sup>261</sup> This was initially made available to users in the EU and India.<sup>262</sup> In addition to recently improving this process, Twitter also boasts a Partner Support Portal (PSP), a tool which allows partners to quickly report suspected violations of Twitter rules, such as misinformation.<sup>263</sup>

In January 2021, Twitter announced a pilot project called “Birdwatch”, where users are able to submit notes on tweets they believe are misleading and identify exactly what they believe is the case in question.<sup>264</sup> The pilot project has seen 1,000 participants, with about 3,300 notes submitted in the first month.<sup>265</sup> Features like this have been experimented with for a while at Twitter.<sup>266</sup>

### Facebook

Since 2016, Facebook has an option for the community to report disinformation. They state: “We’re testing several ways to make it easier to report a hoax if you see one on Facebook, which you can do by clicking the upper right hand corner of a post. We’ve relied heavily on our community for help on this issue, and this can help us detect more fake news.”<sup>267</sup> Their platform currently gives ‘false information’ as one option users can select when reporting content.

### Google

Google has feedback options for its search engine and other products/elements, which allow users to flag content (including as dis/misinformation).<sup>268</sup> Users have to write out their complaint and have the option of including a screenshot.<sup>269</sup> YouTube also allows its users to report content. On any video, you can choose to report them from a dropdown menu, where you can then choose the option of “spam or misleading”.<sup>270</sup>

### Microsoft

LinkedIn allows you to report posts or comments as misinformation.<sup>271</sup>

<sup>261</sup> Twitter Safety, “Strengthening our approach to deliberate attempts to mislead voters,” *Twitter*, April 24, 2019, [https://blog.twitter.com/en\\_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html](https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html).

<sup>262</sup> Ibid.

<sup>263</sup> Twitter, *Twitter Progress Report: Code of Practice on Disinformation*, (Twitter: 2020), 18.

<sup>264</sup> Coleman, “Introducing Birdwatch.”

<sup>265</sup> Tiffany, “Who Would Volunteer to Fact-Check Twitter?”

<sup>266</sup> In 2020, a leaked demo showed Twitter was experimenting with adding bright orange labels in response to community reporting of misinformation. Ben Collins, “Twitter is testing new ways to fight misinformation — including a community-based points system,” *NBC*, February 20, 2020, <https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931>.

<sup>267</sup> Mosseri, “Addressing Hoaxes and Fake News”; Facebook, “Working to Stop Misinformation and False News,” accessed May 11, 2021, <https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news>.

<sup>268</sup> Google, *How Google Fights Disinformation* (Google: 2019), 5.

<sup>269</sup> Google, “Help improve Google’s products,” accessed on May 11, 2021, <https://www.google.com/tools/feedback/intl/en/>; Alex Murray, “How to report fake news to social media,” *BBC*, November 22, 2016, <https://www.bbc.com/news/38053324>.

<sup>270</sup> YouTube Help, “Report inappropriate content,” *YouTube*, accessed on May 11, 2021, <https://support.google.com/youtube/answer/2802027#zippy=>.

<sup>271</sup> LinkedIn Help, “Recognize and Report Spam, Inappropriate, and Abusive Content,” *LinkedIn*, last updated April 2021, <https://www.linkedin.com/help/linkedin/answer/37822>.

**TikTok**

TikTok allows members of the community to report accounts, videos, or comments.<sup>272</sup>

**Other**

**Reddit** is unique, as all of its content is moderated by volunteers from the community (Admins). In order to support these efforts, Reddit has a dedicated subreddit for support.<sup>273</sup> This subreddit often contains posts from reddit employees who provide moderators with updates on what Reddit is doing to tackle misinformation, as well as resources.<sup>274</sup> Admins can also make use of an AutoModerator tool (AutoMod) which can algorithmically remove content, including suspected mis/disinformation.<sup>275</sup> Additionally, the design of Reddit as a platform makes it very open to community moderation: any post can be up or downvoted, and users also have the ability to up or downvote comments on posts. While such features can sometimes enable disinformation to spread within motivated communities, for many of the popular subreddits these features do work to call attention to false claims and news.<sup>276</sup>

Finally, **Apple/Apple News/iTunes** allows users to report any concerns about content, including fake news.<sup>277</sup>

**All platforms have some way for users to report content: however, much like labels, the design and extent of these feature vary across platforms and regions.** Typically, to report, the user receives a number of prompts which the platform uses to categorize their complaint about a piece of content. Some, but not all, have a dedicated option for false information or disinformation (Table 8). There is also a discrepancy in features per region: for instance, Twitter's Birdwatch pilot was only launched in the USA, and Twitter's already existing function to report misleading voting tweets was, as of 2019, only available in Europe and India.<sup>278</sup>

Table 8: Platform Reporting

Platform	Reporting
<b>LinkedIn (Microsoft)</b>	"Suspicious or Fake", which then allows the user to further select "Misinformation."
<b>Facebook</b>	"False Information"
<b>Twitter</b>	[No option to report mis- or disinformation, user will have to fill in a general inquiry]
<b>Google</b>	[No option to report mis- or disinformation, user will have to fill in a general inquiry]
<b>TikTok</b>	[No option to report mis- or disinformation, user will have to fill in a general inquiry]

272 Joe Cornell, "How to Report Videos, Accounts, and Comments on TikTok," *How-To Geek*, February 24, 2020, <https://www.howtogeek.com/658518/how-to-report-videos-accounts-and-comments-on-tiktok/#:~:text=Tap%20the%20three%2Ddot%20icon,a%20description%20of%20your%20report>.

273 Reddit, "Mod Support," accessed May 11, 2021, <https://www.reddit.com/r/ModSupport/>.

274 u/worstnerd (Reddit Admin: Safety), "Misinformation and COVID-19: What Reddit is Doing," *Reddit*, 2020, [https://www.reddit.com/r/ModSupport/comments/g21ub7/misinformation\\_and\\_covid19\\_what\\_reddit\\_is\\_doing/](https://www.reddit.com/r/ModSupport/comments/g21ub7/misinformation_and_covid19_what_reddit_is_doing/).

275 Spandana Singh and Koustubh Bagchi, "How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19: Reddit," *New America*, June 1, 2020, <https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/>.

276 Robin Cohen, Karyn Moffatt, Amira Ghenai, Andy Yang, Margaret Corwin, Gary Lin, Raymond Zhao, Yipeng Ji, Alexandre Parmentier, Jason P'ng, Wil Tan, and Lachlan Gray, "Addressing Misinformation in Online Social Networks: Diverse Platforms and the Potential of Multiagent Trust Modeling," *Information* 11 (2020): 5.

277 Nitasha Tiku, "Why Snapchat And Apple Don't Have A Fake News Problem," *Buzzfeed*, December 1, 2016, <https://www.buzzfeednews.com/article/nitashatiku/snapchat-fake-news>.

278 Twitter Safety. "Strengthening our approach to deliberate attempts to mislead voters."

The lack of a dedicated 'disinformation' reporting option limits the comprehensiveness of the statistics platforms can collect and share on user-reported disinformation on their platform.

Finally, a less-known concern about community reporting is who and how these reports are processed. As mentioned in a 2018 Council of Europe study, "There are strong suggestions that the complete response systems of internet platforms such as Facebook, Google or Microsoft to user queries are automated for many types of inquiries and complaints... Often, many users will need to complain about a specific type of content *before* an automated algorithm identifies it as relevant to be referred to a human operator for content review."<sup>279</sup> While this is intended as a cost-saving measure by platforms, it must not impede both the quality of review for submitted complaints, as well as the appeal process.

**Some have concerns about how useful community reporting is due to concerns over the level of participation by the average user.** This lack of engagement on reporting may stem from the perspective by many platforms, and their users, that the platform is just a service they use rather than a collective good that both users and the platform have a 'collective responsibility' to maintain. This is highlighted when contrasting how Wikipedia operates to other platforms. At its core, Wikipedia relies completely on user-generated content, which is both moderated and created by largely anonymous users. While this may sound ripe for the creation and spread of false information, this is largely not the case. Most users who casually reference Wikipedia understand that the content is created by users, which gives them more hesitancy about trusting Wikipedia blindly – a kind of critical thinking specifically about Wikipedia taught today from elementary schools onward. However, Wikipedia is understood by its users to rely on their contributions, thus cultivating an actively collaborative and community-centered mindset incentivizing actions such as adding and correcting information. As the creator of Wikipedia wrote in 2016, "... the most exceptional thing about Wikipedia is that it is a collaboration, built by volunteer contributors from every walk of life. They come together, from different backgrounds and different beliefs, to learn and understand, to document the past for the future."<sup>280</sup> Similarly, Reddit can be considered another example where users engage collectively in both the creation and moderation of content within their specific 'subreddits' for the incentive of maintaining a positive community.

As such, some note that community moderation attempts by traditional platforms suffer as users often do not see the value of engaging in such projects, especially when platforms like Facebook and Twitter are seen as corporations rather than distinct communities. This is highlighted, for instance, in the recent Twitter Birdwatch pilot. As mentioned earlier, over 1000 people were taking part in this pilot, but together they only produced 3300 labels. This leads observers like Kaitlyn Tiffany to note that "In the Birdwatch pilot's first three weeks, during which Twitter approved its first 1,000 participants, only 3,300 notes were submitted, many of them tests. The company will have to inspire a lot more people to be significantly more invested if this is ever to become a useful backstop against viral misinformation. Doing that will require a major shift on the platform."<sup>281</sup> She questions if there is even a correct motivation for Twitter users to spend time on this feature: somewhat pessimistically, she notes that "Wikipedia was a collectivist project from the very beginning, but Twitter is a for-profit company that operates what is often called a 'data-mining operation' or a 'hell site' by its own

<sup>279</sup> Council of Europe, *Algorithms and Human Rights*, 24-25.

<sup>280</sup> Jimmy Wales, "Wikipedia's strength is in collaboration – as we've proved over 15 years," *The Guardian*, January 15, 2016, <https://www.theguardian.com/commentisfree/2016/jan/15/wikipedia-israel-palestine-15-years-encyclopedia>.

<sup>281</sup> Tiffany, "Who Would Volunteer to Fact-Check Twitter?"



most dedicated users.”<sup>282</sup> Creating vivid community participation may require platforms to encourage users to view a forum like Facebook or Twitter as more of a collective community, where everyone has some role in preventing the spread of disinformation.

**Platforms should offer more transparency into the effectiveness and usage of community reporting against disinformation, providing full datasets and statistics.** In particular, up-to-date data about the effectiveness of community reporting is very difficult to find. In the most recent set of reports about Covid-19 disinformation submitted by platforms to the European Commission, only Twitter really talked about community reporting.<sup>283</sup> The older October 2019 annual self-assessment reports<sup>284</sup> submitted by platforms to the EC as part of the EU Code of Practice on Disinformation are only slightly more illuminating. In their submission, Twitter reports that between May 1-May 20, 2019, 28,456 reports were received for misleading voting related content using their community reporting functions,<sup>285</sup> providing some insight into the feature. Facebook also discusses their reporting features briefly in these reports, but does not provide any statistics.<sup>286</sup> However, there is no real distinction between user-flagged or algorithm-flagged statistics, further undermining these already limited numbers. Given the 2020 reports are not yet released, and the emergence of new Codes such as the Australian Code of Practice on Disinformation and Misinformation (which also heavily feature community reporting as part of their terms<sup>287</sup>), an optimist might hope more transparency will come in the coming few months – however, there is no guarantee this will happen.

Perhaps a reason why there is so little transparency on these statistics is because they do not paint a nice picture for platforms. As mentioned earlier, many observers do believe platforms underutilize community feedback. Other recent reports openly question how often platforms act upon user-generated reporting of misinformation. Notably, a recent report by the Center for Countering Digital Hate (CCDH) found that platforms failed to remove 95% of anti-vaccination misinformation reported to them.<sup>288</sup> Further research by the CCDH also found that platforms did not act upon three quarters of the misinformation reported to them in June 2020.<sup>289</sup>

**Some platforms suggest that their internal processes and algorithms are far quicker than relying on community reporting.** This is highlighted in the numbers Facebook included within their 2019 Facebook Baseline Report on Implementation of the Code of Practice on Disinformation. They state: “In Q2 and Q3 2018, we found and flagged 99.6% of the accounts we subsequently took action on before users reported them. We acted on the other 0.4% because users reported them first. This number increased from 98.5% in Q1 2018.”<sup>290</sup> Overall, this is technically a good thing: ideally, platforms would be able to spot and take down disinformation before consumers can report it. However, no platform beyond Facebook provided such statistics, so it is difficult to make wide conclusions.

<sup>282</sup> Ibid.

<sup>283</sup> And then Twitter only really brought it up to boast about their new Birdwatch pilot project, not supplying any real information or statistics on previous reporting features. Twitter, *Twitter Report: Staying safe and informed on Twitter during COVID-19* (Twitter: 2021).

<sup>284</sup> Despite these being labelled as ‘annual’, I have not found the 2020 reports, should they exist.

<sup>285</sup> Twitter, *Twitter Progress Report*, 17.

<sup>286</sup> Facebook, *Facebook report on the implementation of the Code of Practice for Disinformation* (Facebook: 2019), 18-19.

<sup>287</sup> Digital Industry Group Inc, *Australian Code of Practice on Disinformation and Misinformation* (DIGI: 2021).

<sup>288</sup> CCDH, *Failure to Act*, 11.

<sup>289</sup> Ibid., 12.

<sup>290</sup> Facebook, *Facebook Baseline Report on Implementation of the Code of Practice on Disinformation* (Facebook: 2019), 4.

Three observations can be made resulting from these statements. First, these statistics indicate that algorithms are faster and more reliable tools than relying purely on community reporting. Yet, second, this does not mean there is no role for community reporting: user queries are valuable to discovering new disinformation which may have been missed or to uncover blind spots. However, third, we have to question whether these internal, algorithmic processes are more effective than user-driven community reporting either *by nature* or *by design*. We have seen that platforms often undervalue community reporting, preferring to focus on algorithms as their primary tool to counter disinformation – but, if you focus resources and attention heavily on algorithms and vastly prefer to act on these internal processes and systems, then community reporting will clearly perform worse. This is problematic: the issue would now not be that users do not make enough use of reporting features, but instead that platforms do not dedicate enough resources to these functions being effective. This concern should be ward off through greater transparency into when, why, and how algorithms are used to address disinformation.

### **Factcheckers and factchecking have unclear connections to community reporting.**

As we will explore in greater depth in the next section, factchecking is a valuable part of responding to disinformation for many (but not all) platforms. Factcheckers are involved in the community reporting process: as platforms like Facebook mention, their factcheckers can respond to posts that users have flagged.<sup>291</sup> However, in some cases, the experimenting of platforms with new methods of community reporting appears to be a way to supersede or replace factcheckers. Simply put, platforms have noticed that many of their users do not enjoy the presence of factcheckers: so some have been experimenting with new community reporting or moderation methods as a manner to replace the work factcheckers traditionally do. Twitter, who has traditionally been less supportive of third-party factchecking, proposes one solution: the Birdwatch pilot (mentioned earlier). In their announcement of the program, they specifically said that “people valued notes being in the community’s voice (rather than that of Twitter or a central authority) and appreciated that notes provided useful context to help them better understand and evaluate a Tweet (rather than focusing on labeling content as ‘true’ or ‘false’).”<sup>292</sup> This would, in theory, allow Twitter to rely on its userbase to perform the services the factchecking community would serve on other platforms. As the program is only in its pilot phase, it is difficult to assess the positive or negative impacts of this, but it is a notable development.

291 Facebook, “How Our Fact-Checking Program Works,” *Facebook Journalism Project*, August 11, 2020, <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works>.

292 Coleman, “Introducing Birdwatch.”

## 8.3 Third-Party Factcheckers

### Twitter

Unlike their peers, Twitter does not mention the use of third-party factcheckers. Documents like their progress reports for the Code of Practice on Disinformation<sup>293</sup> are absent of such terms, unlike companies like Facebook and Google, who often emphasize their collaboration with such services. Instead, Twitter's factchecking occurs via internal, closed processes.<sup>294</sup>

### Facebook

Facebook's work with third-party factcheckers runs in conjunction with their own algorithms and services. Third-party factcheckers can identify and rate sources.<sup>295</sup> Then, "If the fact-checking organizations identify a story as false, it will get flagged as disputed and there will be a link to a corresponding article explaining why. Stories that have been disputed also appear lower in News Feed."<sup>296</sup> These factcheckers are all approved by the International Fact-Checking Network.<sup>297</sup> Facebook also publicly discloses the criteria their factcheckers use.<sup>298</sup> In total, Facebook partners with over 80 factchecking organizations globally, which also extend to Covid-19 misinformation.<sup>299</sup>

Facebook outlines a 3-step process to the usage of factcheckers within their content moderation process:<sup>300</sup>

Step	Name	Action
1	Identify	Machine learning/algorithms, community reporting, and factchecker's own observations identify suspicious content.
2	Review	Factcheckers review the content, and apply a variety of labels (False, Altered, Partly False, Missing Context; see Labelling section). They can also report to Facebook the content as True or Satire.
3	Act	If the content is flagged as False, Altered, Partly False, or Missing Context, Facebook takes a variety of actions (Reduced Distribution, Sharing Warning, Sharing Notifications, Misinformation Labels, Removing Incentives for Repeat Offenders).

### Google

Google partners with a number of outside factcheckers and news organizations to combat disinformation and promote quality journalism.<sup>301</sup> As mentioned earlier, Google's fact-checking primarily happens through Schema.org's ClaimReview markup, where websites can submit factchecks. Google then algorithmically reviews whether or not to include a website's factchecks based on a variety of factors..<sup>302</sup>

### Microsoft

Microsoft has a long-standing partnership with NewsGuard, a factchecking organization which now also does work on the Covid-19 crisis. NewsGuard offers a browser plug-in which warns the user if a webpage has potential disinformation (this is especially compatible with Windows browsers) and publishes factchecks.<sup>303</sup>

<sup>293</sup> Twitter, *Twitter Progress Report*.

<sup>294</sup> Elizabeth Culliford and Katie Paul, "With fact-checks, Twitter takes on a new kind of task," *Reuters*, May 31, 2020, <https://www.reuters.com/article/us-twitter-factcheck-idUSKBN2360UO>.

<sup>295</sup> Facebook, "How our Fact-Checking Program Works."

<sup>296</sup> Facebook, "Working to Stop Misinformation and False News."

<sup>297</sup> Facebook, "Partnering with Third-Party Fact-Checkers," *Facebook Journalism Project*, March 23, 2020, <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners>.

<sup>298</sup> Facebook, "How our Fact-Checking Program Works."

<sup>299</sup> Facebook, *Facebook response to the European Commission Communication on Covid-19 Disinformation*, 4.

<sup>300</sup> Facebook, "How our Fact-Checking Program Works."

<sup>301</sup> Google, *How Google Fights Disinformation*, 6.

<sup>302</sup> Google, "Fact Check."

<sup>303</sup> Google, *How Google Fights Disinformation*, 8.

## TikTok

TikTok has stated that they aim to continue expanding work with third-party factcheckers of content, specifically mentioning their growing partnerships with AFP, Animal Politico, Estadão Verifica, Facta.news, Lead Stories, Newtral, Politifact, SciVerify and Teyit.<sup>304</sup> These factcheckers can flag content, which then follows a pre-established internal escalation path that “ensures prompt action.”<sup>305</sup> However, this program only exists in “eight markets” that are left unspecified in TikTok’s 2020 transparency report.<sup>306</sup>

Similarly, the exact steps in this escalation path for factcheckers are not easy to find. For instance, an October 2020 announcement about their Fact-Checking program in the Asia-Pacific market only stated that it “leverages a team of fact-checkers who review and verify reported content. Once the information is confirmed to be false or misleading, we take proactive steps to remove the content in line with our Community Guidelines and notify the user accordingly.”<sup>307</sup>

**Most, but not all, platforms use third-party factcheckers; often claiming it to be highly effective. Little data or transparency supports these claims.** While platforms themselves have often praised the role of third-party factcheckers: i.e. Facebook themselves have said that “Fact-checking is highly effective in fighting misinformation: when something is rated “false” by a fact-checker, we’re able to reduce future impressions of that content by an average of 80%.”<sup>308</sup> However, others have noted that they have “...refused, however, to publicly release any data to support these claims.”<sup>309</sup> So, at the moment, we tend to have statements from platforms which proclaim the effectiveness of third-party factchecking, but no access to data to independently verify such claims.

This leaves us in a precarious situation. We know, from many studies, that third-party fact-checking and labelling can reduce the spread and belief in disinformation.<sup>310</sup> However, these studies are almost all based on data collected from laboratory studies. Consider a 2018 study by James Thorne and Andreas Vlachos examining automated factchecking: they had a number of datasets, but these were relatively small and none had any input from platforms. This was a key limitation they themselves noted, writing “There are currently a limited number of published datasets resources for factchecking.”<sup>311</sup> Increased transparency is clearly needed to adequately assess the role third-party factcheckers should have in fighting disinformation on platforms: something which initiatives like the EU Code of Practice call for, yet have not yet successfully delivered.

304 TikTok, *December 2020 Report EU Code of Practice on Disinformation / COVID-19*, 3.

305 Ibid.

306 TikTok, “TikTok Transparency Report 2020 H1,” September 22, 2020, <https://www.tiktok.com/safety/resources/transparency-report-2020-1?lang=en>.

307 Arjun Narayan Bettadapur, “TikTok partners with fact-checking experts to combat misinformation,” *TikTok*, October 1, 2020, <https://newsroom.tiktok.com/en-au/tiktok-partners-with-fact-checking-experts-to-combat-misinformation>.

308 Meredith Carden, “Responding to The Guardian: A Fact-Check on Fact-Checking,” *Facebook*, December 13, 2018, <https://about.fb.com/news/2018/12/guardian-fact-check/>.

309 Sam Levin, “‘They don’t care’: Facebook factchecking in disarray as journalists push to cut ties,” *The Guardian*, December 13, 2018, <https://www.theguardian.com/technology/2018/dec/13/they-dont-care-facebook-fact-checking-in-disarray-as-journalists-push-to-cut-ties>.

310 Drutman, “Fact-Checking Misinformation Can Work. But It Might Not Be Enough.”; Chan et al, “Debunking.”; Walter et al, “Fact-Checking.”

311 Thorne and Vlachos, “Automated Fact Checking,” 3351.

**There are a variety of common critiques levied against factcheckers.** There are three common critiques we will examine here: bias, low popular support, and who decides what can be factchecked.

**Rather oxymoronically, a number of critics are concerned about the bias of third-party factcheckers, and are calling to ‘factcheck the factcheckers’.** In a March 2021 Wall Street Journal editorial, one such critic voices fear that “Facebook’s [third-party] fact-checkers ‘cherry-pick’... studies to support their own opinions, which they present as fact.”<sup>312</sup> They particularly draw from examples of research on Covid-19, where they are protesting the labeling of scientific research which runs against popular consensus as misinformation. This is by no means an outlier: this line of argumentation is especially common these days, with many accusing factcheckers of being biased towards one political view or another.

**(Relatively) Low popular support.** Skepticism of factcheckers and factchecking on platforms is often seen in the general population. In 2016, a report found that “just 29% of all Likely U.S. Voters trust media fact-checking of candidates’ comments.”<sup>313</sup> A more recent 2019 study found that attitudes had shifted slightly, noting “Overall, Americans are split in their views of fact-checkers: Half say fact-checking efforts by news outlets and other organizations tend to deal fairly with all sides, while about the same portion (48%) say they tend to favor one side.”<sup>314</sup> This is not great for either factcheckers nor platforms. If platforms use factcheckers, a significant percentage of their userbase will be unhappy. Factcheckers themselves need legitimacy to work effectively: something that is becoming more and more difficult in increasingly polarized societies.

**Who decides what can be factchecked?** While platforms like Facebook prolifically hire factcheckers, there are limits to what these factcheckers can do. For instance, commonly, opinion pieces are not eligible for factchecking labels. Yet, the issue then arises in why such limitations exists – an issue which has often been a critique levied at platforms by their hired third-party factcheckers. Facebook has seen its fair share of controversy over such issues. For instance, on Facebook, “... opinion is generally not eligible for fact-check labels, fact-checkers can still label op-eds and similar content if they contain misinformation.”<sup>315</sup> However, third-party factcheckers often push back, and argue that “Facebook still decides what counts as opinion, and it can compel changes to fact-check labels or remove misinformation strikes from a page accordingly.”<sup>316</sup> Given that Facebook retains this final authority, and not the third-party factcheckers, one can question how much power Facebook truly gives their factcheckers in controversial applications. As Andrew Dessler, a factchecker and scientist, noted on this topic, “The most important thing about the story, and something that doesn’t seem to bother a lot of people, is that we have outsourced decisions like this to corporations... This is a truly terrible situation to be in.”<sup>317</sup>

312 The Editorial Board, “Fact-Checking Facebook’s Fact Checkers.”

313 No Author, “Voters Don’t Trust Media Fact-Checking,” *Rasmussen Reports*, September 30, 2016, [https://www.rasmussenreports.com/public\\_content/politics/general\\_politics/september\\_2016/voters\\_don\\_t\\_trust\\_media\\_fact\\_checking](https://www.rasmussenreports.com/public_content/politics/general_politics/september_2016/voters_don_t_trust_media_fact_checking).

314 Mason Walker and Jeffery Gottfried, “Republicans far more likely than Democrats to say fact-checkers tend to favor one side,” *Pew Research Center*, June 27, 2019, <https://www.pewresearch.org/fact-tank/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/>.

315 Alex Pasternack, “Facebook quietly pressured its fact-checkers over climate and abortion posts,” *Fast Company*, August 20, 2020, <https://www.fastcompany.com/90538655/facebook-is-quietly-pressuring-its-independent-fact-checkers-to-change-their-rulings>.

316 Ibid.

317 Ibid.

This feud over what the role of factcheckers should be is more evident in some platforms than others. Most notably, Facebook has had several public falling-out's with a number of major third-party factcheckers, such as in 2019 when both the Associated Press and Snopes (two of the top third-party factcheckers) announced they were cutting ties with Facebook, citing the lack of transparency from Facebook and how the company was too controlling over their factchecking.<sup>318</sup> Snopes employees particularly critiqued the alignment between Facebook's mission and the mission of factcheckers, noting that "Facebook has one mission and fact-checking websites should have a completely different mission."<sup>319</sup> Finally, while Facebook has claimed to make quarterly reports with feedback from factcheckers, third-party factcheckers themselves have claimed not to have seen or received them. Facebook responded to these allegations with the by-now familiar rhetoric stressing the importance they see in their collaborations with third-party factcheckers.<sup>320</sup>

As a sidenote, this debate exposes the difference in how factcheckers view themselves versus how platforms view them. Given the typically limited roles for factcheckers in platform documents, platforms seem to want factcheckers to only provide factchecks for content. Yet, factcheckers themselves are expanding their vision of what they do, now often focusing on making systemic recommendations to address the root causes of disinformation on platforms.<sup>321</sup>

**While platforms do not have any industry best practices for the use of factcheckers, there are industry standards for factcheckers themselves.** As mentioned previously, many do have concerns about what happens if a factchecker is wrong, or their power is abused. To this extent, the factchecking industry has an International Fact-Checking Network (ICFN) and a code of principles.<sup>322</sup> The ICFN in particular is concerned about accountability for their signatories: in the past, they have done investigations into accusations against signatory factchecking organizations for work done for platforms like Facebook.<sup>323</sup> In addition, a variety of other initiatives, such as the American Press Institute's Fact-Checking and Accountability Journalism Project, aim to research, understand, and improve factchecking practices.<sup>324</sup>

However, many have noted that there are no real patterns or best standards when it comes to the use of factcheckers by platforms. As written by the European Commission about the use of factcheckers, "In general, the level and form of such collaboration vary considerably across platforms and Member States, as does the platforms' follow-up with respect to content that has been fact-checked."<sup>325</sup> This is problematic – establishing industry best practices for the use of factcheckers can both make factcheckers more consistent across platforms, as well as minimize any editorial roles the platform may wish to have over the decisions of their factcheckers. Yet, such a development seems rather unlikely: new industry-created initiatives

318 Dave Lee, "Key fact-checkers stop working with Facebook," BBC, February 2, 2019, <https://www.bbc.com/news/technology-47098021>.

319 Levin, "'They don't care': Facebook factchecking in disarray as journalists push to cut ties."

320 Carden, "Responding to The Guardian: A Fact-Check on Fact-Checking."

321 This can be seen in a developing distinction between first- and second-generation factcheckers. For instance, Full Fact, an independent factchecking organization, has distinguished between first and second generation factcheckers. For them, the first generation of factcheckers are those offering services to news organizations to independently verify content and ensure accountability. However, an emerging second generation of factcheckers claim that this first generation does not go far enough, and seek to also focus on action and advocating systemic change. Africa Check, Chequeado, and Full Fact, "Fact checking doesn't work (the way you think it does)," *Full Fact*, June 20, 2019, <https://fullfact.org/blog/2019/jun/how-fact-checking-works/>.

322 Poynter, "The commitments of the code of principles."

323 International Fact-Checking Network, "ICFN releases a statement about accusations against one of its verified signatories," *Poynter*, September 11, 2019, <https://www.poynter.org/fact-checking/2019/ifcn-releases-a-statement-about-accusations-against-one-of-its-verified-signatories/>.

324 Credibility Coalition, "American Press Institute Fact-Checking and Accountability Journalism Project," accessed on May 12, 2021, <https://credibilitycoalition.org/credcatalog/project/american-press-institute-fact-checking-and-accountability-journalism-project/>.

325 EC, *Assessment of the Code of Practice on Disinformation*, 11.



such as the recent Australian Code of Practice on Misinformation and Disinformation support and urge the use of factcheckers, but do not say anything about the recommended usages, application, or oversight.<sup>326</sup>

The work of the ICFN should be commended and encouraged, as it provides at least some level of standardization in the conduct of factcheckers. Platforms should be encouraged to develop and agree to best standards for the use of factcheckers. Recent Code of Practice initiatives should go further than simply recommending the use of factcheckers, but also elaborate on how, when, and why they should be used.

## 8.4 Oversight Boards

### Facebook

As expressed on their website, “The Oversight Board was created to help Facebook answer some of the most difficult questions around freedom of expression online: what to take down, what to leave up, and why.”<sup>327</sup> This independent review board,<sup>328</sup> first operational in 2020, is the first of its kind. It looks at “the most consequential content decisions” made by Facebook and Instagram and issues binding verdicts that can reverse or uphold individual moderation decisions, as well as offer non-binding policy recommendations. Facebook does not have to oblige to the latter, but has 90 days to respond “constructively and in good faith.”<sup>329</sup>

The board is governed by its charter,<sup>330</sup> rulebook,<sup>331</sup> and most importantly its bylaws.<sup>332</sup> It consists of 19 members (will be expanded to 40) with different professional background (legal experts, human rights advocates and journalists) and regional backgrounds (less than half of the board is from the US, Canada and Europe) that each serve a three-year term, up to a maximum of two terms.

The board selects its cases in two ways: it can take on appeals submitted by aggrieved users if they meet the four conditions<sup>333</sup> or through a referral of a “significant and difficult” case from Facebook. The Board is not obliged to take on a case, unless it is submitted as an expedited referral from Facebook for exceptional cases that need to be completed within 30 days.<sup>334</sup> The scope of the Board has been narrowed down to only review content that the company has taken down, therefore skewing the scope to over-moderation by the company, leaving out any of the cases in which they failed to take moderation steps.

<sup>326</sup> DIGI, *Australian Code of Practice on Disinformation and Misinformation*.

<sup>327</sup> Oversight Board, accessed May 12, 2021, <https://oversightboard.com/>.

<sup>328</sup> The Board is a “non-charitable purpose trust” that is part of the Facebook corporation with a contribution of \$130 million from Facebook for the next six years. The trust is run by six trustees selected by Facebook and a corporate trustee that handle the finances.

<sup>329</sup> Nick Clegg, “Welcoming the Oversight Board,” *Facebook*, May 6, 2020, <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.

<sup>330</sup> Facebook, *Oversight Board Charter*, (Facebook: 2019), [https://about.fb.com/wp-content/uploads/2019/09/oversight\\_board\\_charter.pdf](https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf).

<sup>331</sup> Oversight Board, *Rulebook for Case Review and Policy Guidance* (Oversight Board: 2020), <https://oversight-board.com/sr/rulebook-for-case-review-and-policy-guidance>.

<sup>332</sup> Oversight Board, *Oversight Board Bylaws*, (Oversight Board: 2021), <https://www.oversightboard.com/sr/governance/bylaws>.

<sup>333</sup> This includes: (1) an active account (2) a previous verdict from Facebook or Instagram, much like usual appeal courts (3) a submission to the board within 15 days of the verdict (4) content that is not “unlawful in a jurisdiction with a connection to the content”. See Bylaws Ibid.

<sup>334</sup> The expedited review referral is intended for exceptional cases where “content could result in urgent real-world consequences”, which the board must take on and deliver a decision within 30 days. See Bylaws Ibid.

**Facebook  
(cont.)**

This board has judged cases that involve the spread of misinformation with a judgement being given to a French instance of Coronavirus misinformation being included within the Board's first set of ruling on five cases in January 2021.<sup>335</sup>

Finally, Facebook has also created the Data Transparency Advisory Group (DTAG) in 2018.<sup>336</sup> As described by Facebook, "This is an independent body made up of international experts in measurement, statistics, criminology and governance. Their task was to provide an independent, public assessment of whether the metrics we share in the Community Standards Enforcement Report provide accurate and meaningful measures of Facebook's content moderation challenges and our work to address them."<sup>337</sup> This group is led by the Justice Collaboratory at Yale Law School,<sup>338</sup> who have released a first report of recommendations, some of which were adopted.<sup>339</sup>

**TikTok**

On March 2, 2021, TikTok announced the creation of a "Safety Advisory Council" for Europe, specifically to help the company out with content moderation in an environment which will likely see increased obligations on platforms via the Digital Services Act (DSA) in the near future.<sup>340</sup> TikTok reported "The Council will bring together leaders from academia and civil society from all around Europe. Each member brings a different, fresh perspective on the challenges we face and members will provide subject matter expertise as they advise on our content moderation policies and practices. Not only will they support us in developing forward-looking policies that address the challenges we face today, they will also help us to identify emerging issues that affect TikTok and our community in the future."<sup>341</sup>

Two things should be noted about this initiative: first, unlike the Oversight Board, it is not clear whether this Council will have any formal powers at all other than advisory. Second, TikTok did not specifically mention disinformation as an issue for this Council, yet considering its general mandate of content moderation, it seems likely that future elaboration on the focus of the Council will cover disinformation more explicitly.

**Others**

n/a

**The Oversight Board is a groundbreaking experiment in introducing accountability, mediation and transparency to Facebook's content moderation, but given its infancy, it is difficult to judge its effectiveness.** Whenever a measure seeks to limit content permitted on a service, there is always a group of people claiming that it impedes their freedom of expression, or that it lacks legitimacy as soon as it is done outside of the government. This also applies to the Board: Some civil rights advocates were quick to criticize this as merely a PR stunt that draws attention on questions concerning accountability away from the Facebook brand and

<sup>335</sup> Adi Robertson, "Facebook Oversight Board overturns hate speech and pandemic misinformation takedowns," *The Verge*, January 28, 2021, <https://www.theverge.com/2021/1/28/22254155/facebook-oversight-board-first-rulings-coronavirus-misinformation-hate-speech>.

<sup>336</sup> Radha Iyengar Plumb, "An Independent Report on How We Measure Content Moderation," *Facebook*, May 23, 2019, <https://about.fb.com/news/2019/05/dtag-report/>.

<sup>337</sup> Ibid.

<sup>338</sup> No Author, "Justice Collaboratory to Lead Facebook Data Transparency Advisory Group," *Yale Law School*, October 2, 2018, <https://law.yale.edu/yls-today/news/justice-collaboratory-lead-facebook-data-transparency-advisory-group>.

<sup>339</sup> Plumb, "An Independent Report on How We Measure Content Moderation."

<sup>340</sup> Laura Kayali, "TikTok launches 'Safety Advisory Council' in Europe," *Politico*, March 2, 2021, <https://www.politico.eu/article/tiktok-launches-safety-advisory-council-in-europe/>.

<sup>341</sup> Julie de Baillencourt, "Meet TikTok's European Safety Advisory Council," *TikTok*, April 21, 2021, <https://newsroom.tiktok.com/en-eu/meet-tiktok-european-safety-advisory-council-eu>.

criticizing that “it cannot and should not aspire to replace democratic public institutions such as the judiciary.”<sup>342</sup> But one has to bear in mind what signal is sent towards these platforms when embarking on such initiatives.

First, beyond Zuckerberg’s altruistic messaging on the Oversight Board,<sup>343</sup> it is likely the Board is conceived as proof to governments that platforms can sufficiently self-regulate themselves when it comes to content moderation (and, subsequently, disinformation). But the Board is in no position to replace actual courts – after all, it has no legal mandate. The point is that even without a legal mandate, it provides a degree of independent third-party oversight over a company that would otherwise make these content moderation decisions completely on its own. It does not preclude any form of additional regulation.

Second, the Oversight Board’s first decisions showed that the Board is not afraid to hold Facebook accountable by reversing their decisions, and to call for ambitious and systemic reforms to Facebook’s content policy and moderation.<sup>344</sup> It is still too early in the game to determine whether it will be a successful initiative, but overall success will likely depend on the Board’s ability to deal with thorny issues and above all Facebook’s willingness to accept and implement their policy recommendations. One of those thorny issues is what happens when Facebook’s standards are in conflict with international human rights law.<sup>345</sup> In terms of Facebook willingness to implement policy changes, they are not off to a great start: it released an overview of the actions they took in response to the first ruling, highlighting 11 distinct actions.<sup>346</sup> However, critics are mixed on this response: while they agree that these steps are moving in the right direction, they still think that many of the actions Facebook committed to were either already happening, “extraordinarily vague” or simply “illusory”.<sup>347</sup> One of the contributing factors to the mixed response from Facebook to the Board’s recommendations might have to do with the latter’s claim that it is not considering the difficulty of implementing or operationalizing its decisions. While the Board has many respected legal and human rights experts, it is low on content moderation experts. This can lead to many of its recommendations simply overshooting to an extent that they cannot be enforced, potentially undermining the Board’s authority in the long term.<sup>348</sup> Alex Stamos, former CSO of Facebook, explained that “...most of the big content moderation problems for Facebook are not the kind of ‘angels amplifying hate speech on a pin’ problems that this Board is well equipped to deal with. The most pervasive issues are related to Facebook having to make a choice between catching all of the bad stuff and how much over-censorship happens once they have made a decision to take something down. I don’t see a Board that is staffed with legal scholars and that meets part-time as helping with that issue.”<sup>349</sup>

342 Javier Pallero, *Protecting Free Expression in the Era of Online Content Moderation: Access Now’s preliminary recommendations on content moderation and Facebook’s planned oversight board* (AccessNow: 2019), 9.

343 Mark Zuckerberg, “Facebook’s commitment to the Oversight Board,” Facebook, 2019, <https://about.fb.com/wp-content/uploads/2019/09/letter-from-mark-zuckerberg-on-oversight-board-charter.pdf>.

344 Evelyn Douek, “The Facebook Oversight Board’s First Decisions: Ambitious, and Perhaps Impractical,” *Lawfare*, January 28, 2021, <https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical>.

345 Douek, “The Facebook Oversight Board’s First Decisions.”

346 Nick Clegg, “Facebook’s Response to the Oversight Board’s First Set of Recommendations,” Facebook, February 25, 2021, <https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations/>.

347 Evelyn Douek, “The Oversight Board Moment You Should’ve Been Waiting For: Facebook Responds to the First Set of Decisions,” *Lawfare*, February 26, 2021, <https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-facebook-responds-first-set-decisions>.

348 Douek, “The Facebook Oversight Board’s First Decisions.”

349 Alex Stamos, “Alex Stamos talks about Facebook’s Oversight Board,” *Galley by CJR*, 2020, <https://galley.cjr.org/public/conversations/-M74eLMfvkdKpIjRfo4>.

If Facebook seriously considers the recommendations from the Oversight Board, then it should be viewed as a key example of an effective self-regulatory measure. However, if Facebook continues its muddled responses, such as in February 2021,<sup>350</sup> then the Oversight Board should demonstrate to policymakers that it is an ineffective mechanism and that Facebook will continue to ignore any recommendations by optional self-regulatory regimes. In the meantime, it offers some much-needed transparency on Facebook's content moderation tools and processes, which the Board is eager to apply to other platforms.<sup>351</sup>

## 8.5 Community Guidelines

### Twitter

Twitter has developed explicit guidelines to combat disinformation within its community rules that are **coherent with the overall policy framework** of the company. Within the authenticity section of **Twitter Rules**, the company establishes a series of rules on **Civic integrity**, banning to "use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process."<sup>352</sup> These rules are part of a more comprehensive policy that explicitly presents a detailed categorization of different types of misinformation associated to civic processes, encompassing (1) "misleading information about how to participate in an election or other civic process", (2) "misleading information with the intent to intimidate or dissuade individuals from participating in an election or other civic process", (3) "misleading information that is intended to undermine the public's confidence in an election or other civic process", and (4) "false or misleading affiliation."<sup>353</sup>

Regarding the **enforcement mechanisms** of these community rules, the consequences of violating the Civic integrity policy depend on the severity and type of non-compliance, as well as the history of non-compliance of the offending account. In the case that an account has repeatedly violated this policy, Twitter will use a warning system to assess whether other compliance control actions should be considered. This is thought to act as a further counter-measure to the spread of misleading information on the platform. The measures that the company can adopt encompass tweet removals, profile modifications, tags and permanent account suspension and locks.<sup>354</sup>

350 Clegg, "Facebook's Response to the Oversight Board's First Set of Recommendations."

351 Board members already have aspirations beyond just Facebook, as reported in February 2021: "Oversight Board co-chair and former Prime Minister of Denmark Helle Thorning-Schmidt painted a more expansive vision for the group that could go beyond making policy decisions for Facebook. The board co-chair said that if the project proves to be a success, "other platforms and other tech companies are more than welcome to join and be part of the oversight that we will be able to provide." Taylor Hatmaker, "Facebook Oversight Board says other social networks 'welcome to join' if project succeeds," *Techcrunch*, February 11, 2021, <https://techcrunch.com/2021/02/11/facebook-oversight-board-other-social-networks-beyond-facebook/>.

352 Twitter Help Center, "The Twitter Rules," *Twitter*, accessed on May 12, 2021, <https://help.twitter.com/en/rules-and-policies/twitter-rules>.

353 Twitter Help Center, "Civic Integrity Policy," *Twitter*, January 2021, <https://help.twitter.com/es/rules-and-policies/election-integrity-policy>.

354 Ibid.

## Facebook

Even though the company explicitly addresses different forms of disinformation within their Community Standards, the Oversight Board considers Facebook's existing rules on the issue of misinformation to be ambiguous. The Board emphasized that the company's **incoherent policy framework on misinformation** posed a barrier for users to know what content is prohibited.<sup>355</sup>

Facebook follows a *Remove-Reduce-Inform* policy to combat false news: (1) *removing* accounts and content that breach the company's community standards or ad policies, (2) *reduce* the spread of false news and inauthentic content, such as clickbait, and (3) *inform* users by providing context on the posts they see.<sup>356</sup>

Despite the existence of a policy strategy to combat false news, the company's community standards, not only do not provide an explicit definition of misinformation underpinning the overall policy framework but is at times contradicting with the latter.

According to **Facebook's community standards**, "Reducing the spread of **false news** on Facebook is a responsibility that we take seriously. ... [But] There is also a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, significantly reduce its distribution by showing it lower in the News Feed." This, as stated by the company, does not mean Facebook does not take action against disinformation: "Although false news does not violate our Community Standards, it often violates our policies in other categories, such as spam, hate speech or fake accounts, which we remove. For example, if we find a Facebook Page pretending to be run by Americans that's operating out of Macedonia, that violates our requirement that people use their real identities and not impersonate others. So, we'll take down that whole Page, immediately eliminating any posts they made that might have been false."<sup>357</sup>

In addition, Facebook also acts against pages that repeatedly spread disinformation: "Pages and websites that repeatedly share misinformation rated False or Altered will have some restrictions, including having their distribution reduced. They may also have their ability to monetize and advertise removed, and their ability to register as a news Page removed for a given time period."<sup>358</sup>

Another relevant area in the fight against misinformation within the company's community standards is **inauthentic behavior**. As stated in the community standards, the company commits to preventing people from **misrepresenting themselves** on Facebook, using fake accounts, artificially boosting the popularity of content or engaging in behaviors designed to enable other violations under the company's Community Standards<sup>359</sup>

<sup>355</sup> Oversight Board, "Case Decision 2020-006-FB-FBR."

<sup>356</sup> Tessa Lyons, "Hard Questions: What's Facebook's Strategy for Stopping False News?" *Facebook*, May 23, 2018, <https://about.fb.com/news/2018/05/hard-questions-false-news/>.

<sup>357</sup> Ibid.

<sup>358</sup> Facebook Business Help Center, "Fact-Checking on Facebook," *Facebook*, accessed on May 12, 2021, <https://www.facebook.com/business/help/2593586717571940?id=673052479947730>.

<sup>359</sup> Facebook, "Inauthentic Behaviour."

## Facebook (cont.)

Following criticisms on the inconsistency between the company's overall policy and community standards, the Board highlighted that while changes in Facebook's policy are announced in the Newsroom, these changes are often not stated in the Community Guidelines. In line with this criticism and after consideration of a case of health-misinformation associated with COVID-19, the Board has called for the creation of new community standards on health-related misinformation as well as the consolidation of existing rules in place that provide a clear definition of misinformation. The Board also proposed initiatives that would increase the **transparency** on how the company will manage health-related misinformation issues, among others it recommended Facebook to publish a transparency report on the application of Community Standards during the COVID-19 pandemic.<sup>360</sup>

## Google

Google's policy strategy to combat misinformation and disinformation is based on two pillars: (1) *counteract malicious actors*, and (2) *provide users with context*. Their strategy towards combatting disinformation is not only limited to their advertising platforms, but as stated by the company, it "focuses on misrepresentative or harmful behavior by advertisers or publishers while avoiding judgments on the veracity of statements about politics or current events." Thus, while their strategy, does not "classify content as 'disinformation', the company highlights the existence of a series of content policies aimed at preventing deceptive or low-quality content in their platforms, such as the management of 'scrapped' or unoriginal content, misinterpretation, inappropriate content, political influence operations, and election integrity.<sup>361</sup>

Despite these policies, the company's **Community Guidelines** do not explicitly mention disinformation, misinformation, false news or inauthentic behavior. Instead, Google's Community Guidelines present several norms that are applicable in some form to disinformation, namely, norms on spam, deceptive practices, scams, impersonation, hate, and harassment.<sup>362</sup>

Nonetheless, within its basic rules of conducts presented in Google's **Terms of Service**, the company highlights that users must not "abuse or **harm others** or yourself (or threaten or encourage such abuse or harm) - for example, by **misleading**, defrauding, defaming, bullying, harassing or stalking others."<sup>363</sup>

Regarding the **action taken in the case of violation** of these norms, the company "reserves the right to suspend or terminate your access to the services or delete your Google Account if ... [they] reasonably believe that your conduct causes harm or liability to a user, third party, or Google – for example, by hacking, phishing, harassing, spamming, **misleading others**, or scraping content that doesn't belong to you."<sup>364</sup>

Furthermore, within **Google Help Communities Content Policy**, the company prohibits content that involves **impersonation**: "We don't allow impersonation of other people or companies or other behavior that is misleading, deceptive, or fraudulent."<sup>365</sup>

<sup>360</sup> Oversight Board, "Case Decision 2020-006-FB-FBR."

<sup>361</sup> Google, *How Google Fights Disinformation*, 26-28.

<sup>362</sup> YouTube, "Community Guidelines," accessed May 12, 2021, <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#community-guidelines>.

<sup>363</sup> Google, "Privacy and Terms," March 31, 2020, <https://policies.google.com/terms?hl=en>.

<sup>364</sup> Ibid.

<sup>365</sup> Google, "Google Help Communities Content Policy," accessed May 12, 2021, <https://support.google.com/communities/answer/7425194?hl=en>.



## Google (cont.)

The **Covid-19 infodemic** illustrates well the moves Google makes to address disinformation within their policies across their services. They created an explicit **Covid-19 medical misinformation policy** as part of **YouTube's Community Rules** which prohibits content that disputes advice from the World Health Organization (WHO) or local health authorities.<sup>366</sup> This was updated to include claims that contradict expert opinions about vaccines. Google also expanded their advertising policies regarding medical-related misinformation given the global health pandemic. Other temporary measures were also introduced, such as a temporary ban on advertising for masks.<sup>367</sup>

Therefore, while the company has a policy strategy in place to combat misinformation and disinformation, these efforts are still in progress as the absence of definitions of these concepts in its Terms of Service and Community Guidelines demonstrates. Furthermore, the company's policy sets a narrow scope for the issues of misinformation and disinformation only tackling COVID-19 medical misinformation and disinformation on Google's advertising products.

## Microsoft

In 2019, Microsoft wrote that "All of Microsoft's services that display advertising have adopted and vigorously enforce policies prohibiting disinformation."<sup>368</sup> They are active in addressing this, stating "...if Microsoft Advertising becomes aware that an ad suspected of violating its policies is being served to our publishers—for instance, because someone has flagged that ad to our customer support team—the offending ad is promptly reviewed and, if it violates our policies, taken down."<sup>369</sup> These policies are most relevant to LinkedIn and Bing. This commitment against disinformation was seen during the Covid-19 infodemic, when LinkedIn updated various policies which further clarified the type of content and behavior allowed on its platform.<sup>370</sup>

Despite these commitments to combatting disinformation, Microsoft's **Community Code of Conduct** does not explicitly refer to the banning of disinformation activities. It does warn, however, of the immediate banning from its forum in the event of "Impersonating a Microsoft employee, agent, manager, host, administrator, moderator, another user, MVP, or any other person through any means."<sup>371</sup> Further reference to **impersonation**, as a form of misinformation, can be found as a section of the company's **Code of Conduct within their Service Agreement**: "Don't engage in any activity that is fraudulent, false or misleading (e.g., asking for money under false pretenses, **impersonating someone else**, manipulating the Services to increase playcount, or affect rankings, ratings or comments) or libelous or defamatory."<sup>372</sup> In the case of violation of the stated terms, the company warns of its sole discretion to stop providing services or close the account of the person breaching the terms.<sup>373</sup>

366 Google, *EU & COVID-19 Disinformation Google Report*, January 2021 (Google: 2021), 3; YouTube, "Community Guidelines."

367 Google, *EU & COVID-19 Disinformation Google Report*, January 2021, 21.

368 Microsoft, *Microsoft Self-Assessment and Report on Compliance with the EU Code of Practice on Disinformation* (Microsoft: 2019), 3.

369 Ibid., 6.

370 Microsoft, *January Update on Microsoft Corporation's Efforts to Tackle COVID-19 Disinformation* (Microsoft: 2021), 9.

371 <https://answers.microsoft.com/en-us/page/codeofconduct> Microsoft, "Microsoft Community Code of Conduct," accessed May 12, 2021, <https://answers.microsoft.com/en-us/page/codeofconduct>.

372 Microsoft, "Microsoft Services Agreement," August 1, 2020, <https://www.microsoft.com/en/servicesagreement/>.

373 Ibid.

## TikTok

TikTok's **community guidelines** explicitly ban misinformation on their service.

"Misinformation is defined as content that is inaccurate or false. While we encourage our community to have respectful conversations about subjects that matter to them, we do not permit misinformation that causes harm to individuals, our community, or the larger public regardless of intent."<sup>374</sup> Within this definition the community guidelines present different types of misinformation: (1) "misinformation that incites hate or prejudice, (2) "misinformation related to emergencies that induces panic", (3) "medical misinformation that can cause harm to an individual's physical health, (4) "content that misleads community members about elections or other civic processes, (5) "conspiratorial content that attacks a specific protected group or includes a violent call to action, or denies a violent or tragic event occurred" and (4) "Digital Forgeries (Synthetic Media or Manipulated Media) that misleads users by distorting the truth of events and cause harm to the subject of the video, other persons, or society".<sup>375</sup> Furthermore, under the definition of misinformation the company also includes to "engage in coordinated inauthentic behaviors such as the creation of accounts to exert influence and sway public opinion while misleading individuals and our community about the account's identity, location, or purpose."<sup>376</sup>

Furthermore, added to the company's policy banning disinformation, in August 2020, the company announced two major updates on their guidelines trying to improve their content moderation efforts. First, they added a new policy prohibiting synthetic or manipulated content that could mislead users by distorting the truth of events in a way that could harm them. The purpose of this policy was to protect users from shallow or deep fakes. The company highlighted that despite this dimension being already covered in their community guidelines, the update made the policy clearer for users. Secondly, while the company already prohibited content from disinformation campaigns, it made a new policy banning coordinated inauthentic behavior: "Do not engage in coordinated inauthentic activities (such as the creation of accounts) to exert influence and sway public opinion while misleading individuals, our community or the larger public about the account's identity, location or purpose."<sup>377</sup>

The company's enforcement mechanism strategy, heavily embedded in the company's Community Guidelines and Terms of Service, is based on a combination of automated and human content moderation. While the company uses algorithmic models to detect inauthentic behavior, patterns and accounts dedicated to spreading misleading or spammy content, the company's content moderation team responds to emerging trends or threats in collaboration with factchecking partners to verify if content is false or misleading.<sup>378</sup>

Overall, there seems to be a very consistent link between the overall company's strategy on disinformation and its implementation, as shown by the explicit reference to different types of disinformation in the Community Guidelines and the Terms of Service, and the updates made to these documents for them to account for increasingly relevant forms of disinformation, such as synthetic content or inauthentic behavior.

<sup>374</sup> TikTok, "Community Guidelines," Last updated December 2020, <https://www.tiktok.com/community-guidelines?lang=en>.

<sup>375</sup> Ibid.

<sup>376</sup> Ibid.

<sup>377</sup> Vanessa Pappas, "Combating misinformation and election interference on TikTok," *TikTok*, August 5, 2020, <https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok>.

<sup>378</sup> Ibid.

**Community guidelines should be the one-stop overview for all policies, including on disinformation.** Currently, most of the community guidelines, which functions as the main justification for the removal of content from their services, only make limited or even no reference to disinformation. Instead their policy is made up of a patchwork of documents, of which the status is often unknown, making it extremely difficult for the user and oversight bodies to control the platform's policy on disinformation. Platform policies need to be reflected in their community guidelines, allowing us to see how platforms define disinformation and what content falls under that umbrella.

**Policies and platform practices often lack transparency in regards to appeals processes for removed content.** A 2019 report by the Electronic Frontier Foundation uncovered some particularly unappealing trends about platform powers and their lack of transparency. For instance, some platforms, most notably Facebook and Snapchat, do not even provide an appeals process or mechanism for removed or suspended content/accounts.<sup>379</sup> In addition, only one major platform (Reddit) was found to provide 'appeals transparency' – i.e. a regular publication outlining appeals and their outcomes.<sup>380</sup> For any removal mechanism to be just, there should be an option for users to appeal the decision, and receive further elaboration or a reversal.

## 8.6 Algorithmic and automated content moderation

### Twitter

Twitter uses algorithmic processes to recommend content to users. Twitter has not disclosed exactly how this algorithm works, but marketers tend to believe that they use machine learning to sort content on ranking signals, which include **recency, relevance, engagement, rich media, and other factors**.<sup>381</sup> Yet, the nature of this algorithm (recommending similar content, amplifying viral content) has been criticized for potentially enabling the spread of extreme political rhetoric.<sup>382</sup>

As the COVID-19 crisis developed, Twitter applied its machine learning algorithms “to **detect the spread of false information that could harm and flag that content for removal**”. Among these automated content moderation strategies, the company used algorithms to **detect accounts that had been used to deny or advise against adhering to official advice** and had promoted “alternative” treatments proven not to be effective. Algorithms were also programmed to **identify commonly spread falsehoods**, such as the alcohol cure or children being immune to the virus.<sup>383</sup>

379 Gennie Gebhart, “Who Has Your Back? Censorship Edition 2019,” *Electronic Frontier Foundation (EFF)*, June 12, 2019, <https://www.eff.org/wp/who-has-your-back-2019>.

380 Ibid.

381 Katie Sehl, “How the Twitter Algorithm Works in 2020 and How to Make it Work for You,” *Hootsuite*, May 20, 2020, <https://blog.hootsuite.com/twitter-algorithm/#:~:text=Twitter%27s%20algorithm%2C%20like%20most%20social,algorithms%2C%20is%20all%20about%20personalization.&text=All%20social%20algorithms%20use%20machine,rich%20media%2C%20and%20other%20factors>.

382 Oliver Darcy, “How Twitter’s algorithm is amplifying extreme political rhetoric,” *CNN Business*, March 22, 2019, <https://edition.cnn.com/2019/03/22/tech/twitter-algorithm-political-rhetoric/index.html>.

383 Bernard Marr, “Coronavirus Fake News: How Facebook, Twitter, And Instagram Are Tackling The Problem,” *Forbes*, March 27, 2020, <https://www.forbes.com/sites/bernardmarr/2020/03/27/finding-the-truth-about-covid-19-how-facebook-twitter-and-instagram-are-tackling-fake-news/?sh=430a82919771>.

## Facebook

Facebook has released some information on how their news feed operates. They state that “the system determines which posts show up in your News Feed, and in what order, by predicting what you’re most likely to be interested in or engage with. These predictions are based on a variety of factors, including what and whom you’ve followed, liked, or engaged with recently.”<sup>384</sup> Marketers tend to believe that the four most important ranking signals are **relationship, content type, popularity, and recency**.<sup>385</sup> And finally, as mentioned earlier, labels from factcheckers on content also affect the position of content in Facebook’s algorithms.

Nonetheless, given that Facebook’s machine learning algorithms were designed to maximize engagement, ‘toxic’ posts that escape content-moderation filters, will continue to be pushed up in the news feed and will reach a larger audience. In fact, in a public note outlining Facebook’s plans for content moderation, Zuckerberg highlighted the harmful consequences of the company’s engagement strategy by showing how the more likely a post was to violate the platform’s community standards, the more user engagement it received, as the algorithms that maximize engagement reward inflammatory content.<sup>386</sup>

In order to fight the intrinsic vulnerability of its engagement strategy to disinformation, Facebook is using machine learning to fight fake news. First, Facebook is using machine learning to **detect false stories and duplicates of these**.<sup>387</sup> Second, machine learning is being used to **screen metadata on published images and check background information against the context in which they are used**. This allows the company to tackle the use of genuine content, in this case photos, in a fake or misleading context. Third, machine learning algorithms are being used to **identify the origin of false claims**. Thanks to Facebook filters, it is possible to redirect which pages are more likely to share false content based on the profile of page administrators, the behavior of the page, and its geographical location.<sup>388</sup> Finally, in 2019, the company introduced a **metric named Click-Gap which is used by Facebook’s News Feed algorithms** to determine the ranking of a post. It allows the company to limit the spread of websites that generate disproportionate engagement when compared with the rest of the web. If Facebook finds that many links to a certain website are appearing on the platform despite few websites on the broader web being linked to that specific site, it limits the website’s reach.<sup>389</sup>

Instagram, owned by Facebook, uses algorithms to **identify and track hashtags**

**that are frequently used in posts containing false or misleading information.** Furthermore, during the pandemic they started redirecting users searching information on COVID-19 to verified and authoritative information sources. The reason for this was to reduce the number of users sharing unverified information on the social platform.<sup>390</sup>

384 Akos Lada, Meihong Wang and Tak Yan, “How does News Feed predict what you want to see?” *Facebook*, January 26, 2021, <https://tech.fb.com/news-feed-ranking/>.

385 Paige Cooper, “How the Facebook Algorithm Works in 2021 and How to Make it Work for You,” *Hootsuite*, February 10, 2021, <https://blog.hootsuite.com/facebook-algorithm/>.

386 Karen Hao, “How Facebook got addicted to spreading misinformation,” *MIT Technology Review*, March 11, 2021, <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>.

387 Mark Zuckerberg, “A Blueprint for Content Governance and Enforcement,” *Facebook*, May 5, 2021, <https://www.facebook.com/notes/751449002072082/>.

388 No Author, “Facebook using machine learning to fight fake news,” *Internet of Business*, accessed May 12, 2021, <https://internetofbusiness.com/facebook-machine-learning-fake-news/>.

389 Emily Dreyfus and Issie Lapowsky, “Facebook Is Changing News Feed (Again) to Stop Fake News,” *Wired*, April 10, 2019, <https://www.wired.com/story/facebook-click-gap-news-feed-changes/>.

390 Marr, “Coronavirus Fake News.”

## Google

Google stresses they frequently update their algorithms, often with a specific eye on combatting the spread of disinformation.<sup>391</sup> Google claims to use a **ranking algorithm**, which they say “...elevates the relevant information that our algorithms determine as the most authoritative and trustworthy above information that may be less reliable.”<sup>392</sup> Simply put, algorithms will mark content as disinformation and consequently downrank it.

They also “... continuously invest in the testing and improving of our ranking algorithms – rolling out thousands of updates every year (e.g., more than 2,400 updates to Google Search algorithms in 2017 alone).”<sup>393</sup> They, for instance, updated their Search Quality Rater Guidelines to enhance flagging by raters on different low-quality webpages, such as misleading information, offensive content, hoaxes and unsupported conspiracy theories. These guidelines allow the company’s algorithms demote this low-quality content and improve the quality of user’s searches over time.<sup>394</sup>

In addition, the company has established direct feedback loops to algorithmic search features, such as Autocomplete and Featured Snippets by including labeled categories so that users can inform the platform directly about whether they find sensitive, inaccurate or offensive content. This feedback has been incorporated into the features allowing to improve user’s search results over time.<sup>395</sup>

Finally, Google also runs a “How Search Algorithms Work” webpage for users and researchers to better understand their search algorithm.<sup>396</sup>

## Microsoft

There are a few relevant excerpts which are fairly illuminative on their own regarding Microsoft’s algorithms: “To combat this type of abuse, Bing Search implements a sophisticated **ranking process** that is increasingly focused not only on ensuring that users always see the **most relevant results** for their search query on the first page, but also that such **results are “high authority”** (unless a user’s query clearly intends to find low authority content). Bing Search is constantly refining its detection algorithms and the metrics it uses to measure them in order to prevent manipulation of its search results by bad actors and to ensure that high-quality sites rank higher than low-quality ones.”<sup>397</sup>

In order to leverage user’s flags that feed their search algorithms efficiently, the company developed a novel algorithm, named DETECTIVE, that **detects false news while jointly learning about users’ flagging accuracy over time**.<sup>398</sup>

Furthermore, Microsoft Advertising uses a **filtration system to detect bot traffic**. This system is embedded in several algorithms that allows it to automatically detect and neutralize invalid or malicious online traffic that can arise from click fraud, fishing, malware, or account compromise.<sup>399</sup>

<sup>391</sup> Google, *How Google Fights Disinformation*, 12.

<sup>392</sup> Ibid., 12.

<sup>393</sup> Google, EC EU Code of Practice on Disinformation: Google Annual Report (Google: 2019), 19.

<sup>394</sup> Ben Gomes, “Our latest quality improvements for Search,” *Google*, April 25, 2017, <https://www.blog.google/products/search/our-latest-quality-improvements-search/>.

<sup>395</sup> Ibid.

<sup>396</sup> Google, “How Search algorithms work,” accessed May 12, 2021, <https://www.google.com/search/howsearch-works/algorithms/>.

<sup>397</sup> Microsoft, *Microsoft Self-Assessment and Report on Compliance with the EU Code of Practice on Disinformation*, 9.

<sup>398</sup> Sebastian Tschitschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. “Fake News Detection in Social Networks via Crowd Signals,” *Companion Proceedings of the The Web Conference 2018* (April 2018), <https://www.microsoft.com/en-us/research/publication/fake-news-detection-social-networks-via-crowd-signals/>.

<sup>399</sup> Ibid., 9.

TikTok	n/a.
Other	<p>In 2017, <b>Snapchat</b> announced updates and new features which may have increased the prominence of disinformation on the platform. In response, they announced that “The Snapchat solution is to rely on algorithms based on your interests — not on the interests of ‘friends’.”<sup>400</sup> Their <b>algorithm thus tries to understand what its users have enjoyed looking at, instead of presenting them with content obtained through feeds by friends or network effects.</b><sup>401</sup></p> <p>It has been argued that this algorithmic model helps guard against false news.<sup>402</sup> Many observers claimed that this was a more straightforward and user-friendly approach than what a lot of other platforms were doing.<sup>403</sup></p>

**Algorithms can enhance the spread and proliferation of disinformation, while lacking the incentives to address the systematic issues which make this possible.** Most platforms heavily rely on algorithms to suggest and find content for users. This often leads to two effects facilitating the spread and proliferation of disinformation: *personalization* and *amplification*.<sup>404</sup>

- *Personalization* entails that platforms use algorithms to scope the interests of the users, and then use that data to present content which aligns with their views.<sup>405</sup> This can lead to ‘filter bubbles’ – when the users are continually presented information which only supports their viewpoints.<sup>406</sup> This can help spread or reinforce disinformation.
- *Amplification* is the use of algorithms to find content users are interested in, and then quickly spread it, thus driving engagement.<sup>407</sup> However, misleading or false content is often accidentally spread due to the similarities they share with actual viral content. Moreover, malicious actors can create false content specifically tailored to abuse these features.

Both these effects are often not addressed sufficiently by platforms, even under self-regulatory regimes like the EU Code of Practice. As critics identify,<sup>408</sup> both personalization and amplification are used by platforms to drive user engagement and interaction, which in turn generates more profit for platforms. Any changes made to these algorithms to mitigate the impact or pervasiveness of these effects will likely result in impacts to a platforms' profits: a sacrifice we have not seen platforms willing to make.

400 Jamie Condliffe, “Snapchat Has a Plan to Fight Fake News: Ripping the ‘Social’ from the ‘Media’,” *MIT Technology Review*, November 29, 2017, <https://www.technologyreview.com/2017/11/29/147413/snapchat-has-a-plan-to-fight-fake-news-ripping-the-social-from-the-media/>; Mike Shields, “Snap suddenly has a leg up on Facebook and Google — but it still needs to do 2 things to steal their advertisers,” *Business Insider*, October 7, 2017, <https://www.businessinsider.com/snapchats-closed-doors-keep-fake-news-out-2017-10?international=true&r=US&IR=T>.

401 Condliffe, “Snapchat Has a Plan to Fight Fake News: Ripping the ‘Social’ from the ‘Media’.”

402 Evan Spiegel, “How Snapchat is separating social from media,” *Axios*, November 29, 2017, <https://www.axios.com/how-snapchat-is-separating-social-from-media-2513315946.html>.

403 Condliffe, “Snapchat Has a Plan to Fight Fake News: Ripping the ‘Social’ from the ‘Media’”; Shields, “Snap suddenly has a leg up on Facebook and Google — but it still needs to do 2 things to steal their advertisers.”

404 DG CNECT, *Study on media literacy and online empowerment issues raised by algorithm-driven media services*, 12.

405 Ibid., 41.

406 Ibid., 41.

407 Ibid., 43-45.

408 See, for instance, Wardle and Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policymaking*, 52.



**Algorithms have often been critiqued for carrying a variety of human rights risks – but the obvious solution (increased transparency of algorithms) has widely been resisted by platforms.** There is much concern about the human rights implications of algorithms. Many point especially towards the impact algorithms can have on freedoms like expression and assembly. Actors like the Council of Europe have supported such concerns, writing in 2018 that “Algorithms are widely used for content filtering and content removal processes [...] directly impacting on the freedom of expression and raising rule of law concerns (questions of legality, legitimacy and proportionality).”<sup>409</sup> The first step towards addressing these concerns is almost unanimously supported by advocates: namely, increasing the transparency of algorithms used by platforms. This transparency should extend to both clarifying the impact of algorithmic processes on user experience, as well as providing researchers access and insight into the code used by algorithms.<sup>410</sup> Yet, despite this pressure, platforms have strongly resisted such efforts – often citing algorithms as “commercially sensitive information” or expressing concerns about such transparency impacting their business model.<sup>411</sup>

**Related to the lack of transparency in algorithms, true accountability for algorithms is also something which does not really currently exist and should be enforced.** Given the hesitancy to allow proper transparency into algorithms, it should not be surprising that there is a lack of processes overseeing algorithms used by platforms. However, some governments are beginning to pass legislature mandating greater transparency of algorithms – although this regulation has largely been completely ineffective.

A prime example is the 2018 *Law on the fight against information disorder* in France, which included a large focus on increasing the algorithmic transparency of large platforms with a particular eye on preventing disinformation.<sup>412</sup> For instance, a 2019 report submitted to their Secretary of State for Digital Affairs, recognizes that “Algorithms are tools that may be misused or misappropriated,”<sup>413</sup> citing a variety of human rights concerns, and also advocating transparency as the solution. In response, this report bluntly states that “Transparency will be effective only if it results from regular dialogue with operators and a process of trial and error mimicking the development process of those algorithms.”<sup>414</sup> Other sources have also recognized that these new French laws put “pressure on platforms to establish a tool for its users to flag disinformation content, establishing more transparency on the workings of their algorithms along with other media literacy initiatives.”<sup>415</sup> Yet, despite these ambitions which followed the recommendations from academics and the civil society, this law has remained relatively ineffective. As reported in a 2019 assessment of the *Law* conducted by the Conseil supérieur de l’audiovisuel (CSA), “the operators’ responses differ very little from the information already accessible on their websites and do not allow the CSA to carry out an exhaustive analysis. Despite the necessary confidentiality of certain information regarding operators’ business models, which are largely based on the performance of their algorithms, the CSA

409 Council of Europe, *Algorithms and Human Rights*, 18.

410 DG CNECT, *Study on media literacy and online empowerment issues raised by algorithm-driven media services*, 59-61.

411 Ibid., 62.

412 France, “Against information manipulation,” November 20, 2018, <https://www.gouvernement.fr/en/against-information-manipulation>; French Secretary of State for Digital Affairs, *Creating a French framework to make social media platforms more accountable: Acting in France with a European vision* (Paris: 2019) <https://thececre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf>.

413 French Secretary of State for Digital Affairs, *Creating a French framework to make social media platforms more accountable*, 25.

414 Ibid., 26.

415 DG CNECT, *Study on media literacy and online empowerment issues raised by algorithm-driven media services*, 56.

is concerned about *the lack of clarity regarding the intelligibility of these algorithms and the incompleteness of the information submitted*.<sup>416</sup> Their recommendations therefore call heavily for increased transparency. However, if platforms were unwilling to supply transparency initially, why would their openness be changed now? More worryingly, this example demonstrates how protective platforms are of their algorithms. This was a French law, legally passed and put into practice, which the platforms, for all intensive purposes, ignored. Clearly, such legal requirements are not enough motivation for platforms to submit to the desired levels of algorithmic transparency.

This result is echoed in other initiatives. The EU Code of Conduct on Disinformation also includes concrete requests for algorithmic transparency: “In line with the HLEG Report and the Communication, the Signatories of this Code acknowledge the importance to ‘take the necessary measures to enable privacy compliant access to data for fact-checking and research activities’ and to ‘cooperate by providing relevant data on the functioning of their services, including data for independent investigation by academic researchers and general information on algorithms.’”<sup>417</sup> Yet, the extent to which major platforms are abiding by this request is questionable: in their 2019 annual self-assessment reports,<sup>418</sup> only Microsoft and Google discussed the operational aspects of algorithms at all – and in a fairly basic manner. The topic of algorithmic transparency was also rather conveniently omitted from the European Commission’s 2020 Assessment of the Code of Practice on Disinformation. Despite calls for algorithmic transparency forming two concrete action points in the original text, these points are not assessed or even really acknowledged in this report, despite platforms clearly not meeting these requirements. More surprisingly, this entire 28-page review only mentions the word “algorithm” twice. To a casual observer, it simply appears that the European Commission all but forgot to look at algorithmic transparency.

416 Conseil Supérieur de l’Audiovisuel (CSA), *Combating the dissemination of false information on online platforms: an evaluation of the application and effectiveness of the measures implemented by operators in 2019* (Paris: CSA, 2020), 7.

417 EC, *EU Code of Practice on Disinformation*, 8.

418 European Commission, “Annual self-assessment reports of signatories to the Code of Practice on Disinformation 2019,” last updated March 8, 2021, <https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019>.

## 8.7 Verified Information Features

### Twitter

Twitter has created “Twitter Event Pages” for 30 countries which bring together the latest information from their respective governments on the Covid-19 crisis.

Twitter has sponsored a series of Twitter Moments and marketing campaigns aimed at spreading Covid-19 awareness.<sup>419</sup> Twitter also has a ‘Curation team’, which finds and highlights “great tweets” as Twitter Moments. This team is specially trained, and currently serves 16 markets in five languages (English, Japanese, Arabic, Spanish and Portuguese).<sup>420</sup>

In addition, Twitter also has, on occasion, attached links and labels to authoritative information sources on certain tweets and accounts. Examples include, during the US 2018 midterm and 2020 elections, labelling accounts of candidates<sup>421</sup>; the labelling of misleading tweets about the coronavirus pandemic and vaccine with links to authoritative information.<sup>422</sup>

### Facebook

Facebook has launched a variety of information centers, which are pages on their platform that contain authoritative information on a topic. Most notable is their Covid-19 Information Center, which is heavily promoted on Facebook and contains information from leading global and national experts and governments.<sup>423</sup> The other topics include a Voting Information Center for the 2020 US elections (United States),<sup>424</sup> and a recently-launched Climate Science Information Center (United States, France, Germany and the United Kingdom, will be expanded later).<sup>425</sup>

419 Twitter, *Twitter Report: Staying safe and informed on Twitter during COVID-19*.

420 Twitter Help Centre, “Twitter Moments guidelines and principles,” *Twitter*, accessed May 12, 2021, <https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles>.

421 Bridget Coyne, “Introducing US Election Labels for Midterm Candidates,” *Twitter*, May 23, 2018, [https://blog.twitter.com/en\\_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html](https://blog.twitter.com/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html); Bridget Coyne, “Helping identify 2020 US election candidates on Twitter,” *Twitter*, December 12, 2019, [https://blog.twitter.com/en\\_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html](https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html).

422 Twitter, *Twitter Report: Staying safe and informed on Twitter during COVID-19*; Musadiq Bidar, “Twitter will label posts with misleading information about COVID-19 vaccines,” *CBS News*, March 2, 2021, <https://www.cbsnews.com/news/twitter-covid-19-vaccine-misinformation-labels/>.

423 Nick Clegg, “Combating COVID-19 Misinformation Across Our Apps,” *Facebook*, March 25, 2020, <https://about.fb.com/news/2020/03/combating-covid-19-misinformation/>.

424 Facebook, “Launching Our Voting Information Center on Facebook and Instagram,” August 13, 2020, <https://www.facebook.com/business/news/launching-our-voting-information-center-on-facebook-and-instagram>.

425 Foo Yun Choo and Katie Paul, “Facebook launches climate science info center amid fake news criticism,” *Reuters*, September 15, 2020, <https://www.reuters.com/article/facebook-climatechange-int-idUSKB-N2660M5>.

## Google

Google has a variety of initiatives which provide their users with authoritative information. They offer a “Fact Check Explorer” tool, which allows a user to enter a term or trending story and see fact-checked news sources.<sup>426</sup>

They have also been rolling out new features in response to infodemics, such as the one associated with the Covid-19 pandemic. Starting in the UK, Google will include information panels on authorized vaccines in your region when one searches for vaccine-related information. This feature is now available in 17 countries, and is being expanded. Searches on Covid-19 and Vaccines now also first display panels with authoritative information. These include data and are simply visualized. Similar panels also show up on YouTube. Google Maps has also been updated to make it easier to find Covid-19 testing sites, masks, health-care options, etc.

YouTube has also been developing informational ‘factcheck’ panels which contain authoritative information which give more context to videos on the platform and are based off the work of third-party factcheckers.<sup>427</sup> These have especially been useful in recent times as YouTube notes more and more people are using YouTube as a source of news and information.<sup>428</sup> YouTube and Google also operate “Breaking News” information/knowledge panels which provide authoritative information on developing and new news stories.<sup>429</sup> Finally, in Google News, there is often the option to view the “Full Coverage” of an event.

## Microsoft

Bing also points “users to special COVID-19 ‘information hubs’ on Microsoft News”<sup>430</sup> using information from trusted news sources. As Microsoft reported, “In December 2020, the Bing COVID experience had 19,378,414 visitors.”<sup>431</sup> This portal also links European users to the EU Covid-19 portal.

LinkedIn, since it is rarely used for disinformation, has instead focused mainly on spreading government information on Covid-19 to local audiences. This includes prompts and mentions in widely distributed newsletters.

Finally, Microsoft has launched the “Microsoft Video Authenticator”, which can “analyze a still photo or video to provide a percentage chance, or confidence score, that the media is artificially manipulated.”<sup>432</sup>

## TikTok

TikTok runs and updates an in-app information hub about vaccines. Users are directed there when they search for vaccine information on the app. These are often market-specific: i.e. for France, there will be informational TikToks created by the French government’s account.

Moreover, TikTok has also began implementing a tag to all content with “Covid-19” terms, where a banner directs a user to an authoritative source.<sup>433</sup>

<sup>426</sup> Google, “Fact Check Explorer,” accessed May 12, 2021, <https://toolbox.google.com/factcheck/explorer>.

<sup>427</sup> YouTube Help, “See fact checks in YouTube search results,” *YouTube*, accessed May 12, 2021, <https://support.google.com/youtube/answer/9229632?hl=en>.

<sup>428</sup> YouTube, “Expanding fact checks on YouTube to the United States,” April 28, 2020, <https://blog.youtube/news-and-events/expanding-fact-checks-on-youtube-to-united-states>.

<sup>429</sup> Google, *How Google Fights Disinformation*, 5; Google, “About knowledge panels,” accessed May 12, 2021, <https://support.google.com/knowledgepanel/answer/9163198?hl=en>.

<sup>430</sup> Microsoft, *January Update on Microsoft Corporation’s Efforts to Tackle COVID-19 Disinformation*, 2.

<sup>431</sup> Ibid., 3.

<sup>432</sup> Burt, “New Steps to Combat Disinformation.”

<sup>433</sup> Kevin Morgan, “Taking action against COVID-19 vaccine misinformation,” *Twitter*, December 15, 2020, <https://newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation>.

**Given the voluntary and platform-driven nature of Verified Information Features, many are the ‘first line of defense’ against disinformation that consumers interact with.** At their core, after all, Verified Information Features provide the consumer with easily accessible, visually appealing, and authoritative information on a subject. Moreover, considering that all major platforms are creating such features, most internet users are being exposed to them: as Google recently estimated, their information panels and features have been viewed 400 billion times.<sup>434</sup> Exposing users to healthy and authoritative information sources is beneficial to protect citizens against disinformation.

Each platform tends to run their own version of a hub for popular issues. For instance, Google, Facebook, and Twitter all have information features for Covid-19:<sup>435</sup> while they provide similar information from similarly authoritative government sources, they do not follow similar design principles. Increasing the coherence among these features may increase the ease of access for users.

These features all tend to link to authoritative, primarily government, sources which are typically clearly labelled as such. There does not appear to be any regulation mandating the creation of such features; instead, most platforms are framing these as voluntary and cooperative actions. This is especially common in the wake of prior infodemics, such as election cycles or the Covid-19 pandemic.

Platforms should be encouraged to further incorporate such Verified Information Features when dealing with disinformation crises.

<sup>434</sup> Google, *EU & COVID-19 Disinformation Google Report*, January 2021, 7.

<sup>435</sup> Facebook, “Coronavirus (COVID-19) Informatiecentrum,” accessed on May 12, 2021, [https://www.facebook.com/coronavirus\\_info/](https://www.facebook.com/coronavirus_info/); Twitter, “Updates on Covid-19 in the Netherlands,” accessed May 12, 2021, <https://twitter.com/i/events/1244645077797851137>; Google, “COVID-19,” accessed on May 12, 2021, <https://www.google.com/search?q=covid+19&oq=covid+19&aqs=edge.0.0j69i60j69i61j69i60j0l3.1874j0j1&sourceid=chrome&ie=UTF-8>.

# 9 Annex II: Platform Governance and Cooperation across the Regime Complex: Case Insights

This Annex describes two case studies to assess how industry cooperation can be facilitated and how compliance with norms can be encouraged. This includes the relationships that platforms have with each other and with government and civil society stakeholders in the context of countering disinformation, as well as the regulatory approaches and cooperative arrangements that can be put in place to advance the industry commitments responsibly. The Global Internet Forum to Counter Terrorism (GIFCT), which is a more well-established initiative that is directly relevant in terms of process and only indirectly relevant in terms of content; and the EU Code of Practice on Disinformation, which is a recent industry-led initiative under the auspices of the EU that is directly relevant both in terms of content and process. The case studies offer a concise description of the initiative including its background and rationale, identify strengths and weaknesses, on the basis of which they arrive at recommendations on how to leverage cooperation and governance mechanisms for the counter-disinformation regime complex and advance the small-n norms listed in Chapter 5 through a coregulation model proposed in Chapter 6.



## 9.1 Case Study 1: The Global Internet Forum to Counter Terrorism (GIFCT)

The first case study takes a closer look at the Global Internet Forum to Counter Terrorism (GIFCT). While it does not deal with disinformation, it does represent an important model for the future of industry cooperation, in particular centralized cooperation, when it comes to online content moderation dealing with harmful content.

### 9.1.1 Background

In the mid-2010s, Internet platforms were increasingly seeing their services being used by terrorist organizations to spread malicious content. Instances like the 2013 Westgate Mall shooting, where terrorists sent over 500 real-time tweets during their attack,<sup>436</sup> and ISIS propaganda shared on Twitter and Facebook throughout the early 2010s<sup>437</sup> massively increased the public and government pressure on platforms to take meaningful actions against online violent extremism and terrorism. Some actions were taken by platforms themselves. Most major platforms quickly updated policies and community standards denying terrorists accounts and posts, and frequently took these down or banned them.<sup>438</sup> Yet, this approach, which lacked any regulatory mechanisms, was not sufficient enough to curb the problem of online terrorist content.

In a cooperative effort to minimize terrorist content, the Netherlands, the UK, Germany, Belgium and Spain sponsored a European Commission Project in 2010 called “Clean IT”, which would develop “general principles and best practices” to combat online terrorist content and “other illegal uses of the internet [...] through a bottom up process where the private sector will be in the lead.”<sup>439</sup> Initially, the initiative, featuring heavy European law enforcement participation, was considering forward-leaning proposals, such as removing anonymity by pushing platforms to enact a real-name policy and to only allow real pictures of the users. This was met with push-back from civil society and ultimately led to the end of the initiative.<sup>440</sup> Nonetheless, experts believe that the “the project helped set the ideological foundations for the European Union’s approach to online terrorist content by advocating for more aggressive terms of service and industry takedowns without formalized legislation.”<sup>441</sup>

From 2013 to 2015, many governments were becoming increasingly frustrated with the slow progress platforms were making on this issue, thus resulting in renewed efforts for a

436 David Mair, “#Westgate: A Case Study: How al-Shabaab used Twitter during an Ongoing Attack,” *Studies in Conflict and Terrorism* 40, no. 1 (2017), <https://doi.org/10.1080/1057610X.2016.1157404>.

437 Julia Greenberg, “Why Facebook and Twitter Can’t Just Wipe Out ISIS Online,” *Wired*, November 21, 2015, <https://www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/>; Home Office (UK) and the Department for Education (UK), *How social media is used to encourage travel to Syria and Iraq: Briefing note for schools* (UK: 2015), [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/440450/How\\_social\\_media\\_is\\_used\\_to\\_encourage\\_travel\\_to\\_Syria\\_and\\_Iraq.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/440450/How_social_media_is_used_to_encourage_travel_to_Syria_and_Iraq.pdf).

438 Nicole Softness, “Terrorist Communications: Are Facebook, Twitter, and Google Responsible for the Islamic State’s Actions?” *Journal of International Affairs* 70, no. 1 (Winter 2016): 205; Jillian C. York, “Terrorists on Twitter: Attempts to silence ISIS online could backfire,” *Slate*, June 25, 2014, <https://slate.com/technology/2014/06/isis-twitter-suspended-how-attempts-to-silence-terrorists-online-could-backfire.html>.

439 Robert Gorwa, “The platform governance triangle: conceptualising the informal regulation of online content,” *Internet Policy Review* 8, no. 2 (June 2019), DOI: 10.14763/2019.2.1407.

440 EDRI, “RIP CleanIT,” January 29, 2013, <https://edri.org/our-work/rip-cleanit/>.

441 Robert Gorwa, “Regulating them softly,” *Centre for International Governance Innovation*, October 28, 2019, <https://www.cigionline.org/articles/regulating-them-softly>.

government-backed initiative tackling this. While the United States held hearings on terrorist content online, little substantive regulation arose out of these sessions.<sup>442</sup> Meanwhile in 2014, the European Commission introduced their plans for an “EU Internet Forum”, which brought together EU member states with big tech platforms including Google, Facebook, Microsoft, and Twitter to discuss how platforms could best counter terrorist content and illegal hate speech.<sup>443</sup> These dialogues eventually led these parties to agree on the EU Code of Conduct on Countering Illegal Hate Speech Online, which committed platforms to swiftly removing hateful or terrorist content, with monitoring provided by a network of civil society organizations.<sup>444</sup> As a result, the platforms updated their terms of service globally and began regular reporting on their progress.<sup>445</sup> Several other platforms joined the initiative in the following years.<sup>446</sup> Overall, the Code was widely considered to be necessary, albeit with concerns about the way it involved civil society in the process and its impact on freedom of speech and expression. More worrying were concerns about the effectiveness of the Code: for instance, one year into the Code, the EC found that Twitter was falling short to meet standards created to remove content. In addition, the EC also reported in 2017 that Facebook and Google only barely met the Code’s standards on ensuring hate speech was removed, with Twitter still falling significantly below the threshold.<sup>447</sup> These uninspiring numbers were noted by European lawmakers, some of which began openly exploring more stringent regulatory schemes, which could include fines for platforms.<sup>448</sup> Only months after these renewed pressures, major platforms announced the creation of a new, industry-driven body which allowed them to address these concerns: the GIFCT.

## 9.1.2 The GIFCT

The Global Internet Forum to Counter Terrorism (GIFCT) was established in 2017 by Twitter, Facebook, Microsoft, and YouTube.<sup>449</sup> This initiative came directly after European governments called for more regulation and requests to companies to both set up an industry forum on this issue and develop new tools to advance the discovery and removal of terrorist content.

The 2019 terrorist attack in Christchurch, broadcasted live on Facebook, marked a watershed moment for the GIFCT. It led to the establishment of the Christchurch Call, a series of voluntary commitments that brings together stakeholders from government, industry and civil

442 <https://www.govinfo.gov/content/pkg/CHRG-114hhrg92852/html/CHRG-114hhrg92852.htm>

443 Kirsten, Fiedler, “EU Internet Forum against terrorist content and hate speech online: Document pool,” *European Digital Rights*, March 10, 2016, <https://edri.org/eu-internet-forum-document-pool/>.

444 European Commission, “The EU Code of conduct on countering illegal hate speech online,” June 30, 2016, [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en); Europa Nu, “The Code of conduct on countering illegal hate speech online,” June 22, 2020, [https://www.europa-nu.nl/id/vl9qfzaji8mu/nieuws/the\\_code\\_of\\_conduct\\_on\\_countering?ctx=vg9pj7ufwbwe&tab=0](https://www.europa-nu.nl/id/vl9qfzaji8mu/nieuws/the_code_of_conduct_on_countering?ctx=vg9pj7ufwbwe&tab=0).

445 EC, “The EU Code of conduct on countering illegal hate speech online.”

446 Ibid.

447 Mark Scott, “Twitter Fails E.U. Standard on Removing Hate Speech Online,” *New York Times*, May 31, 2017, <https://www.nytimes.com/2017/05/31/technology/twitter-facebook-google-europe-hate-speech.html>.

448 Melissa Eddy and Mark Scott, “Facebook and Twitter Could Face Fines in Germany Over Hate Speech Posts,” *New York Times*, March 14, 2017, <https://www.nytimes.com/2017/03/14/technology/germany-hate-speech-facebook-tech.html>.

449 Google, “Featured Policies,” *Google Transparency Report*, accessed on May 11, 2021, <https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism/@policy>; “Update on the Global Internet Forum to Counter Terrorism,” *Twitter*, December 4, 2017, [https://blog.twitter.com/en\\_us/topics/events/2017/GIFCTupdate.html](https://blog.twitter.com/en_us/topics/events/2017/GIFCTupdate.html); Facebook, “Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism,” June 26, 2017, <https://about.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/>; Microsoft Corporate Blogs, “Global Internet Forum to Counter Terrorism has first meeting Aug. 1,” *Microsoft*, July 31, 2017, <https://blogs.microsoft.com/on-the-issues/2017/07/31/global-internet-forum-counter-terrorism-first-meeting-aug-1/>.

society against online violent extremist content. Today, the Call has been signed by more than 50 countries and international organizations, and also includes ten leading tech companies among its signatories. It operationalized three crisis response protocols: The Christchurch Call Shared Online Crisis Response Protocol, the industry-led Content Incident Protocol (CIP), and the European protocol. Taken together, these protocols provide “...an interlinking communications network that enables a rapid and coordinated response to online events, between governments and companies.”<sup>450</sup> The Call was also the driving impetus for reform of the GIFCT: it was no longer simply a coalition of the willing run by industry representatives, but was given an independent non-profit status and dedicated resources, including an executive director, staff, operating boards and an Independent Advisory Committee, in addition to a part to carry out a number of the Christchurch Call's commitments.<sup>451</sup>

In its current state, the GIFCT now includes most social media platforms.<sup>452</sup> To become a member, companies need to show that they prohibit terrorist content in their terms of service and/or community guidelines, have a capacity to act on the basis of these prohibitions, and are willing to explore technological solutions to this end. They will have to report regularly on their performance, respect human rights in content moderation, and commit to include civil society in combatting violent extremism online. The GIFCT also has an internal capacity building mechanism for new members, which is run in cooperation with the Tech Against Terrorism initiative, instructing members on issues such as what an effective terms of service looks like and what content moderation capabilities are available.<sup>453</sup>

The GIFCT houses a number of specific initiatives which combat terrorism. One of these initiatives started a few years before the inauguration of the GIFCT in 2016: the Hash-Sharing Consortium, where known hashes of terrorist videos and content are shared between 13 companies and platforms.<sup>454</sup> Hashing is a digital fingerprinting technology for tagging violative content that is then shared in an anonymized way with partners so they can quickly identify and decide to take appropriate measures against it. It is previously known as PhotoDNA, which was developed to counter online child sexual abuse material (CSAM), and repeatedly re-occurs in other fields dealing with harmful content. Another initiative facilitated by the GIFCT is the Content Incident Protocol (CIP), an established process in which GIFCT members are quickly made aware of potential content from a terrorist event circulating online, allowing them to take quick action against it.<sup>455</sup> This process has been frequently used. Finally, the GIFCT also began a program in 2019 which shares URLs linked to terrorist content amongst platforms, thus enabling coherence in content takedown. Platform centralized

450 New Zealand Foreign Affairs & Trade, “Christchurch Call,” accessed on May 11, 2021, <https://www.mfat.govt.nz/en/peace-rights-and-security/international-security/christchurch-call/>.

451 <https://www.christchurchcall.com/christchurch-call.pdf>; <https://gifct.org/governance/#government>: The GIFCT's operational budget and mission alignment are all annually set and monitored by their Operating Board: “The Operating Board is composed of GIFCT's founding members; at least one rotating company from the broader membership cadre; new companies that meet leadership criteria, and the rotating chair of the Independent Advisory Committee, who participates as a non-voting member. The Operating Board chair rotates annually.” As such, industry, civil society, and government all have voices in the governance and direction of the GIFCT. Christchurch Call, *The Christchurch Call to Action To Eliminate Terrorist and Violent Extremist Content Online*, <https://www.christchurchcall.com/christchurch-call.pdf>; Global Internet Forum to Counter Terrorism, “Governance,” accessed on May 11, 2021, <https://gifct.org/governance/>.

452 Global Internet Forum to Counter Terrorism, “Membership,” accessed on May 11, 2021, <https://gifct.org/membership/>.

453 GIFCT, “Membership”; GIFCT, “Joint Tech Innovation”; Tech Against Terrorism, accessed on May 11, 2021, <https://www.techagainstterrorism.org/>; Jen Patja Howell, “The Lawfare Podcast: Collaborating to Counter Violent Extremism Online,” *Lawfare*, November 25, 2020, <https://www.lawfareblog.com/lawfare-podcast-collaborating-counter-violent-extremism-online>.

454 Global Internet Forum to Counter Terrorism, “What is the hash sharing consortium and how does it work?” accessed May 11, 2021, <https://gifct.org/?faqs=what-is-the-hash-sharing-consortium-and-how-does-it-work>.

455 GIFCT, “Joint Tech Innovation”.

cooperation in the GIFCT took precedence over individual companies developing their own technology to deal with terrorist content and striking favors from regulators and users. These norms, standards and collaboration, rose not only because of a sense of moral obligation: “there are also reputational reasons - it could be advantageous for companies to present a united front and avoid being singled out for difficult choices in areas not considered core to their product.”<sup>456</sup>

### 9.1.3 Analysis and Conclusions

The GIFCT has a number of success it can point to; however, critics still exist with valid concerns. In terms of successes, the GIFCT has been particularly active in facilitating inter-platform responses and capacity building, thereby rightly highlighting information sharing initiatives such as the sharing of hashes, URLs and other content linked to terrorists for removal. Yet, at the same time, it has been criticized for its lack of transparency that gives rise to due process, accountability and freedom of expression concerns. In this analysis, several lessons will be extracted from this model, in particular concerning transparency and accountability, legitimacy, standards, capacity building, and taxonomies.

**As the ambiguity and dubiety of a defined threat rises, there are corresponding increases in the proclivity of errors in content moderation, the difficulty of developing technical tools, and, ultimately, a greater risk for a centralized approach.** This is reflected in the GIFCT, especially as it itself still struggles to clearly define the concept of terrorism. This is perhaps not unexpected: although there are UN resolutions, national legislations, foreign terrorist lists, and a large swath of academic work to rely upon, there is no dominant international definition of ‘terrorist content.’ Such uncertainty over definitions has implications, with authoritarian regimes using the loose definition of ‘terrorist content’ to crack down on dissenting or critical voices against their regime.<sup>457</sup> The lack of a dominant definition also affects content moderation: when there is no agreed definition of a particular violation, content moderation becomes difficult. UN Special Rapporteur Kaye notes this issue explicitly when discussing the GIFCT: “This is not like child sexual abuse, for which there is a consensus around imagery that clearly and objectively meets a concrete definition. Rather, it is asking companies to make legal decisions, and fine ones at that, about what constitutes the elements of terrorism, of incitement to terrorism, of the glorification of terrorism.”<sup>458</sup> Companies are aware that unlike child abuse, there are legitimate reasons to share terrorist propaganda, such as for research, journalistic, counterterrorist or other purposes. Yet, many critics still point to the frequency that such content decisions, on what constitutes a violation, takes place through a collective black box process removed from public scrutiny. As some note, this makes it “easier for mistakes to be missed, and for members to shirk blame for any that are found, by making it more difficult to identify the source.”<sup>459</sup>

Evelyn Douek places the tensions between well- and ill-defined taxonomies of threats on a spectrum. On the one end is online child abuse: it has a clear definition and is a threat with a high risk of harm that is universally prohibited. This clarity facilitates centralized industry cooperation and the development of technical mitigation tools. On the other side of the spectrum

<sup>456</sup> Evelyn Douek, “The Rise of Content Cartels,” *Knight First Amendment Institute at Columbia University*, February 11, 2020, <https://knightcolumbia.org/content/the-rise-of-content-cartels>.

<sup>457</sup> Tech Against Terrorism, accessed on May 11, 2021; Global Network on Extremism and Technology, accessed on May 11, 2021, <https://gnet-research.org/>.

<sup>458</sup> David Kaye, *Speech Police: The Struggle to Govern the Internet* (New York: Columbia Global Reports, 2019), 83.

<sup>459</sup> Douek, “The Rise of the Content Cartels.”

are the violations that are difficult to define and not considered illegal *per se*. Disinformation can be considered one of those, especially when it deals with political speech. Without a clear definition, technical tools are hard to develop, content moderation is more error-prone, and centralized cooperation among platforms might seem most attractive because platforms fear the backlash of unilateral action. But Douek warns that platform policies dealing with content should remain their own transparent decision in accordance with their community guidelines, not one made behind an “unaccountable consensus”.<sup>460</sup> She adds that in order to determine whether such a centralized approach is useful for cases in between the two ends of the spectrum, “the answer should depend on an empirical inquiry into factors such as the prevalence of that category of content; the accuracy of the relevant technology; the cost and practicality of small platforms developing similar tools; the relevant risk of harm; and, especially, the contestability of the category definition and whether it implicates speech, such as political speech, that is ordinarily highly protected.”<sup>461</sup> Terrorism is one of these cases in between: it is still more manageable than foreign influencing or disinformation because there are FTO lists and UN sanctioning arrangements on specific terrorist organizations.

**Without meaningful transparency, there can be no accountability for industry-led content moderation initiatives.** Outsiders have criticized that the relatively closed-off “Shared Industry Hash Database” of terrorist propaganda allow these platforms the power to decide what qualifies as terrorist content without much transparency about the location of the database,<sup>462</sup> its contents, and its process (how entries are made and by whom). In addition, there are no independent, public or centralized mechanisms to audit, challenge, or remediate entries into the database.<sup>463</sup> Nor are there provisions for third-party researcher access. It resembles a ‘black box’ where only consortium members know what and how entries are added, by whom, and if there are mechanisms to resolve a dispute.

The GIFCT’s centralized industry-wide hashing means that a piece of content that is classified as a violation by one platform can be removed by the other members, without a centralized remedy mechanism or public oversight function. This means that a bias or error in labeling from one platform is therefore likely to be repeated by the other members. Even if discovered, the affected party cannot rely on the same centralized mechanism to remedy the decision, being instead forced to approach each platform individually. As such, the likelihood of companies duplicating each other’s errors is at best unclear, or at worst considerably high. Douek observes the effects of the voluntary nature of the initiative on transparency: “Platforms try to have it both ways by touting the benefits of collaboration while insisting that inclusion in the GIFCT database does not mean automatic removal by all members, as they all reach “independent” decisions. Without transparency, it is impossible to evaluate these claims. There is no information available about how often members reject other members’ determinations or what happens when this occurs.”<sup>464</sup>

<sup>460</sup> Douek, “The Rise of the Content Cartels.”

<sup>461</sup> Ibid.

<sup>462</sup> The unknown location of the database has been criticized, although it appears to be hosted by one of its member companies. The GIFCT is acutely aware of the policy implications for moving this database to the GIFCT, but is as of now still considering the technical feasibility of creating and hosting that infrastructure.

<sup>463</sup> Chloe Hadavas, “The Future of Free Speech Online May Depend on This Database,” *Slate*, August 13, 2020, <https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html>; Emma Llansó, “Platforms Want Centralized Censorship. That Should Scare You,” *Wired*, April 18, 2019, <https://www.wired.com/story/platforms-centralized-censorship/>.

<sup>464</sup> Douek, “The Rise of the Content Cartels.”

**A centralized approach, like the GIFCT, not only has a risk for unaccountable mistakes, but can simultaneously create an illusion of legitimacy that individual company decisions lack.** In turn, this can provide a shield against scrutiny. It risks embedding and legitimizing certain content moderation standards without much oversight or critical public contestation. When, for example, social media giants de-platformed former US President Donald Trump in January 2021, most suggested an objectivity or even legitimacy by merely referencing each other's decisions without much of a formalized process or external verification.<sup>465</sup> Similarly to this informal ad-hoc decision making, the GIFCT risks a more formalized way of greenwashing that allows companies to simply refer to their membership of the GIFCT or the number of hashes shared when they are enquired or questioned about how they fight the spread of violent extremism or terrorism online. Such references have little value when the GIFCT and the hashes operate as part of a 'black box' with no way of verifying the success of these results. This raises another issue: identifying metrics of success.

**Centralization can enhance the role of already-powerful platforms in deciding the standards and objectives for the smaller players.** Because they have the capacity and resources that allow them to decide what constitutes violent extremism online, they set the standard for others to follow. To this end, enforcement gaps can emerge that favor actions against violations and violators that are well reported on by certain governments over underreported groups.<sup>466</sup> This is seen in the terrorism case study by an overrepresentation of information on Islamic terrorist organizations (thanks to the focus on them by actors like the US government) compared to groups such as white extremists (for which no similar lists are maintained by the US government). Moreover, there also, arguably, exists a bias to focus on US lists; which are then applied and propagated on a global level by the larger platforms and copied by their smaller partners.

**Centralization can facilitate capacity building for more targeted content moderation in favor of protecting online speech.** The GIFCT and even wider collaboration has been relatively successful in capacity building for its smaller members, especially in terms of otherwise expensive content moderation tools dealing with violent extremism.<sup>467</sup> The development of automated content moderation tools at scale is very hard and resource intensive, requiring large datasets that are not available to smaller companies. When content moderation is not straightforward, especially for disinformation, the struggle for smaller companies to do this becomes bigger, and the drive for a centralized approach becomes higher. In a way, collaborative efforts resolve the capacity and resource gap between larger and smaller companies in a similar way as David Kaye, the UN special rapporteur for freedom of expression, has called for in the context of developing tools to detect hate speech: "[t]he largest companies should bear

<sup>465</sup> <https://www.chathamhouse.org/2021/01/deplatforming-trump-puts-big-tech-under-fresh-scrutiny> This lack of respect for formal rules or processes have been noted in the months since this decision: notably, in May 2021, Facebook's Oversight Board noted that while there were legitimate reasons to remove Trump, Facebook issued an indefinite ban – something that is not really covered in their user agreement. So, they found that "It is not permissible for Facebook to keep a user off the platform for an undefined period, with no criteria for when or whether the account will be restored." Simply put, even if removing Trump was the right decision to make, Facebook should still have followed their formalized process – they should not have the authority to validate their decisions just because they are Facebook, or all their competitors are doing so. Harriet Moynihan, "Deplatforming Trump puts big tech under fresh scrutiny," *Chatham House*, January 22, 2021, <https://www.chathamhouse.org/2021/01/deplatforming-trump-puts-big-tech-under-fresh-scrutiny>; Oversight Board, "Oversight Board upholds former President Trump's suspension, finds Facebook failed to impose proper penalty," May 2021, <https://oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/>.

<sup>466</sup> Chris Meserole and Daniel Byman, *Terrorist Definitions and Designations Lists: What Technology Companies Need to Know* (London: RUSI, 2019), 9-13.

<sup>467</sup> Youtube spent over 100\$ million on developing the tech it uses to identify copyright violations. See Google, "How Google Fights Piracy," November 2018, [https://www.blog.google/documents/25/GO806\\_Google\\_FightsPiracy\\_eReader\\_final.pdf](https://www.blog.google/documents/25/GO806_Google_FightsPiracy_eReader_final.pdf).



the burden of these resources and share their knowledge and tools widely, as open source, to ensure that smaller companies, and smaller markets, have access to such technology.”<sup>468</sup> When capacity building is done well, as was the case for some of the GIFCT’s successes, smaller companies are able to take less drastic solutions for identifying and removing terrorist propaganda and thus better protect rights like online speech. It also makes these smaller companies better equipped to resist governmental pressure to remove speech: i.e. they can point to a tested procedure.

In short, centralized approaches should not allow platforms to launder difficult decisions through untransparent and unaccountable centralized processes that create an illusion of legitimacy.

### 9.1.4 Recommendations

Based on the analysis of the GIFCT, the following recommendations are made on centralized industry cooperation that will inform the coregulation and cooperation model for the counter-disinformation regime complex described in Chapter 7.

**Build in organizational and reporting transparency.** To achieve meaningful multistakeholder collaboration, the GIFCT Advisory Board could commit more strongly to organization transparency. This has not escaped civil society leaders: for instance, according to Andrew Sullivan, Present and CEO of the Internet Society, “To achieve this [meaningful multistakeholder collaboration], the GIFCT advisory panel can be made more useful through meaningful and binding commitments to organizational transparency: make the board of the GIFCT work in public and let us all understand what it is doing, and use the advisory panel to supervise that.”<sup>469</sup> In terms of reporting, the question arises as to what constitutes sufficient transparency. The GIFCT’s transparency reports have been notoriously short and reveal very little meaningful information about its operations and processes.<sup>470</sup> At the very minimum, transparency mandates should demand metrics that are auditable and incentivize desirable behaviors (i.e. not just results-oriented, but also focusing on the process).

**Incentivize (right) transparency.** Civil society often criticizes what little transparency large platforms and their centralized collaborations projects provide as simply a PR stunt. While this criticism is often valid, it disincentivizes companies that decide to be transparent, while companies that are silent and do not report at all may potentially fly under the radar. At the same time, there should be an understanding of the notion that, especially in the case of foreign influence operations, transparency standards can actually be counterproductive when it comes to deterring malicious actors, as they could adjust their modus operandi in such a way to circumvent content moderation triggers and effectively game the system.

**Legitimacy through transparency.** The GIFCT has shown that while it can be a forum for effective collaboration for its members, such collaboration also needs to be seen. Without the transparency and accountability mechanisms described above, the so-called benefits or added value of cooperation cannot be verified. Rather than relying on the judgement of individual companies, the content moderation standards and tools ought to be legitimized by

<sup>468</sup> Kaye, *Speech Police*, 83.

<sup>469</sup> Andrew Sullivan, “Looking the GIFCT in the Mouth,” *The Internet Society*, October 11, 2019, <https://www.internetsociety.org/blog/2019/10/looking-the-gifct-in-the-mouth/>.

<sup>470</sup> Global Internet Forum to Counter Terrorism, “Transparency,” accessed on May 11, 2021, <https://gifct.org/transparency/>.

anchoring them in universally accepted norms and due process, including transparency and oversight mechanisms.

**Build in meaningful independent oversight**, embedded into the institutional design with mandates for human rights audits, allowing independent civil society organizations to check the processes and outcomes of the collaboration. Building in third-party oversight and accountability is crucial to avoid abuse by and lack of responsibility from members. Importantly, oversight mechanisms need to be strong enough so red flags can be raised on the basis of sufficient information. Auditing can also occur through sampling rather than on an absolute basis.

**Accountability requires access to remediation.** When a centralized decision-making process is established for censorship, it should be accompanied by a centralized remediation process. This means, for example, when a piece of content is taken down as a result of the GIFCT hash-sharing, then a challenge to this takedown on any individual platform needs to feed back into the GIFCT network so that decisions made collectively do not need to be challenged individually.”<sup>471</sup> Applied in the context of disinformation/influence campaigns, there is a need for more insight into the extent companies rely on the signals of their peers as well, in addition to a remediation process for those who would like to challenge these actions. Such a centralized format would be especially beneficial for smaller platforms that do not necessarily have the resources or the will to create an appeals process.

---

<sup>471</sup> Douek, “The Rise of the Content Cartels.”

## 9.2 Case Study 2: The EU Code of Practice on Disinformation

The EU Code of Practice on Disinformation (hereafter referred to as 'the Code') was not only selected because it directly deals with the issue of disinformation, but also because of its self-regulation method in which corporations, under the clerical oversight function of the European Commission, developed their own commitments when it comes to using content moderation to counter disinformation.

### 9.2.1 Background

The High-Level Expert Group on Fake News and Online Disinformation (HLEG) convened in January 2018, composing of a variety of experts from academia, the private sector, the public sector, and civil society. It was formed to advise the European Commission on how to form a policy to counter fake news and disinformation online. This ultimately resulted in the March 2018 report titled "A multi-dimensional approach to disinformation," that presented 10 guiding principles which the HLEG (including representatives from platforms such as Twitter and Facebook) agreed should form a starting point for the development of an EU Code of Practice on disinformation (see Table 9).<sup>472</sup>

*Table 9: The 10 Guiding Principles of the EU High-Level Expert Group on Fake News and Online Disinformation*

#### The 10 Guiding Principles of the EU High-Level Expert Group on Fake News and Online Disinformation

1. Platforms should adapt their advertising policies, including adhering to "follow-the-money" principle, whilst preventing incentives that lead to disinformation, such as to discourage the dissemination and amplification of disinformation for profit. These policies must be based on clear, transparent, and non-discriminatory criteria;
2. Platforms should ensure transparency and public accountability with regard to the processing of users' data for advertisement placements, with due respect to privacy, freedom of expression and media pluralism;
3. Platforms should ensure that sponsored content, including political advertising, is appropriately distinguished from other content;
4. Platforms should take the necessary measures to enable privacy-compliant access to data for factchecking and research activities;
5. Platforms should make available to their users advanced settings and controls to empower them to customize their online experience;
6. Platforms should, in cooperation with public and private European news outlets, where appropriate take effective measures to improve the visibility of reliable, trustworthy news and facilitate users' access to it;
7. Where appropriate, trending news items should, if technically feasible, be accompanied by related news suggestions;
8. Platforms should, where appropriate, provide user-friendly tools to enable users to link up with trusted factchecking sources and allow users to exercise their right to reply;
9. Platforms that apply flagging and trust systems that rely on users should design safeguards against their abuse by users;
10. Platforms should cooperate by i.e. providing relevant data on the functioning of their services including data for independent investigation by academic researchers and general information on algorithms in order to find a common approach to address the dissemination and amplification of disinformation.

<sup>472</sup> EC, *A multi-dimensional approach to disinformation Report of the independent High level Group on fake news and online disinformation*, 32-33.

The HLEG's principles and proposed plan led to a Communication from the European Commission in April 2018 that set out their main objectives, an action plan, and self-regulatory tools to tackle online disinformation.<sup>473</sup> One of the proposed actions was the establishment EU Multistakeholder Forum on Disinformation to establish a Code of Practice.<sup>474</sup> The Forum was convened in May 2018, consisting of a Working Group and a Sounding Board. The Working Group, consisting of representatives from major online platforms, their trade association, ad exchanges, and major advertising associations, delivered a draft two months later. The Sounding Board, composed of media and civil society organizations, found that the draft required "significant improvements, in terms of concrete commitments and clarity on who commits to what, measurable Key Performance Indicators (KPIs) and redress mechanisms for potential breaches of the Code."<sup>475</sup> At the fourth and final meeting of the Forum in September 2018, the Working Group delivered a final draft. The Sounding Board members were not able to support the draft and published an opinion later that month concluding that "The 'Code of Practice' as presented by the working group contains no common approach, no clear and meaningful commitments, no measurable objectives or KPIs, hence no possibility to monitor process, and no compliance or enforcement tool: it is by no means self-regulation, and therefore the Platforms, despite their efforts, have not delivered a Code of Practice."<sup>476</sup>

## 9.2.2 The Code of Practice

Despite its controversial birth, the EU Code of Practice on Disinformation rose above the concerns of the Sounding Board and became the first self-regulatory set of standards to fight disinformation voluntarily developed by and for platforms, officially signed into effect in October 2018. It includes dedicated commitments on five thematic areas: ad placements, political and issue-based advertising, integrity of services, empowering consumers, and empowering the research community.<sup>477</sup> While the Sounding Board of the Code may reject the self-regulatory nature of the Code because it felt the Code fell short, both academics and the EU recognize the Code as prime example of such an initiative.<sup>478</sup>

At the time of writing, the Code has 16 signatories, including 6 major platforms (Facebook, Google, Twitter, Mozilla, Microsoft, and TikTok) and numerous trade associations.<sup>479</sup> The platforms also have each submitted a roadmap detailing their commitments to the Code and their current actions against disinformation, which can be found on the European Commission's website.<sup>480</sup> The website also hosts several implementation reports by some of the original signatories (including Facebook, Google, Mozilla, and Twitter) which detail the steps taken so far in accordance with the Code.

<sup>473</sup> EC, *Tackling online disinformation*.

<sup>474</sup> European Commission, "Meeting of the Multistakeholder Forum on Disinformation," 2018, <https://digital-strategy.ec.europa.eu/en/library/meeting-multistakeholder-forum-disinformation>.

<sup>475</sup> European Commission, "A draft code of practice on online disinformation," Last updated: March 8, 2021, <https://digital-strategy.ec.europa.eu/en/library/draft-code-practice-online-disinformation>.

<sup>476</sup> The Sounding Board, "The Sounding Board's unanimous final opinion on the so-called code of practice," *EBU*, September 28, 2018, <https://www.ebu.ch/news/2018/09/sounding-board-of-forum-on-disinformation-online-issues-unanimous-opinion-on-so-called-code-of-practice>.

<sup>477</sup> Elaboration on all these commitments can be found in the text of the EU Code of Practice on Disinformation (pages 4-8).

<sup>478</sup> EC, *Assessment of the Code of Practice on Disinformation*; Chris Marsden, Trisha Meyer, and Ian Brown, "Platform values and democratic elections: How can the law regulate digital disinformation?" *Computer Law & Security Review* 36 (2020): 12.

<sup>479</sup> European Commission, "Roadmaps to implement the Code of Practice on disinformation," Last updated: March 8, 2021, <https://digital-strategy.ec.europa.eu/en/news/roadmaps-implement-code-practice-disinformation>.

<sup>480</sup> Ibid.

While the Code is an industry-led initiative, it is part of a wider EU effort against disinformation in which the European Commission has a clerical or watchdog function. Overall, the Code produced mixed results: it has enabled structured cooperation between industry and public policy, but it lacks uniform definitions, procedures, key performance indicators, and has obvious limitations intrinsic to the self-regulatory nature and lack of transparency measures. The European Commission notes that these limitations impact the Code's effectiveness, writing:

"At present, it remains difficult to precisely assess the timeliness, comprehensiveness and impact of platforms' actions, as the Commission and public authorities are still very much reliant on the willingness of platforms to share information and data. The lack of access to data allowing for an independent evaluation of emerging trends and threats posed by online disinformation, as well as the absence of meaningful KPIs to assess the effectiveness of platforms' policies to counter the phenomenon, is a fundamental shortcoming of the current Code."<sup>481</sup>

As a result, strong trust has not been built between industry, governments, academia and civil society.

The Commission therefore suggests a number of improvements for the code – such as "commonly-shared definitions, clearer procedures, more precise commitments and transparent key performance indicators and appropriate monitoring" – as well as calling for further effort to broaden participation, in particular from the advertising sector. It also wants to see a more structured model for cooperation between platforms and the research community. The European Commission has also announced new regulations in development which would address some of these shortcomings or gaps, including the 2020 Digital Services Act and an update to the EU Code itself as announced in the December 2020 European Democracy Action Plan.<sup>482</sup>

### 9.2.3 Analysis and Conclusions

While the Code is generally considered to be a welcome and necessary initiative, it is not without criticism. This analysis is structured along two main questions: First, are the commitments of the EU Code effective at addressing disinformation? And, second, how effective is the regulative approach the EU Code endorses?

In terms of determining the effectiveness of the Code, the European Council's assessment, and an external study that informed its findings, is especially useful.<sup>483</sup> This study examines the impact of the Code on the actions of the platforms and identifies areas for improvement. Let us briefly examine their most pressing findings for each of the five pillars of the Code:

<sup>481</sup> EC, *Assessment of the Code of Practice on Disinformation*.

<sup>482</sup> The Digital Services Act (DSA) was introduced in December 2020 by the European Commission. While it does not directly deal with disinformation, this act introduces new rules and regulations for platforms, some of which are aimed specifically at combatting illegal content and promoting transparency. The European Democracy Action Plan clearly outlines the vision for the DSA, stating: "The Digital Services Act (DSA) will propose a horizontal framework for regulatory oversight, accountability and transparency of the online space in response to the emerging risks. It will propose rules to ensure greater accountability on how platforms moderate content, on advertising and on algorithmic processes." Simply, the DSA is envisioned to emerge as a co-regulatory backstop to the soon-to-be updated EU Code of Practice on Disinformation. European Commission, "FAQ — Digital Services Act," last updated: December 15, 2020, <https://ec.europa.eu/digital-single-market/en/faq/faq-digital-services-act>.

<sup>483</sup> Iva Plasilova, Jordan Hill, Malin Carlberg, Marion Goubet, and Richard Procee, *Study for the "Assessment of the implementation of the Code of Practice on Disinformation"* (Luxembourg: Publications Office of the European Union, 2020), <https://digital-strategy.ec.europa.eu/en/library/study-assessment-implementation-code-practice-disinformation>.

ad placements, political and issue-based advertising, integrity of services, empowering consumers, and empowering the research community.

***The impact of the Code on the scrutiny of ad placements was limited: while the platforms already had the respective policies and processes in place to take down ads, many blind spots remained.*** The platforms did include quantifiable numbers of takedowns of ads for violating content. However, many of these appeared to align more with preexisting policies rather than anything mandated by the Code. More worryingly, a number of blind spots and areas for improvement remain. It was especially noted that many platform policies are not aligned with each other and do not share common definitions, thus limiting the coherence of efforts combatting ads. Other blind spots also remained unaddressed. For instance, it was noted that platform policies do not make proper distinctions between ads on their own platforms and advertising on third-party platforms by their services: a clear blind spot in terms of coherence that collaboration should be able to make some progress in. Finally, the platforms seemed much more content to focus on the “low-hanging fruit” of removing “imposter websites” (websites which misrepresent their purpose and thus are easy to tackle within preexisting policies) rather than the more pervasive problems of ad placements on websites which openly convey disinformation.<sup>484</sup>

***The Code had the strongest impact on the transparency of political and issue-based advertising.*** While it remains difficult to trace the platform changes back to the Code, it must be noted that the platforms put in place systems to label or even ban political ads and issue-based advertising.<sup>485</sup> However, there were still issues. Importantly, no signatories provided either information on or tools needed to measure the transparency of these measures. The Code also mandated the creation of online libraries of political ads with Application Programming Interfaces (APIs) that was implemented by signatories but could be much improved in terms of their functionality and completeness. Overall, the Code had a legitimate impact in terms of political advertising on platforms: yet, much progress remains to be made specifically in providing transparency on the steps they did take so third parties can fully measure the extent of any progress made so far.

***The impact of the Code on the integrity of services is limited as it is not possible to determine whether platform tools and policies emerged as a result of the Code.*** While platforms provided information on their efforts to remove fake accounts/bots/spam and disclosed numbers on accounts removed, there was insufficient detail and transparency to adequately measure the effectiveness of the implemented measures.

***The impact of the Code on the empowerment of consumers remains limited because of the lack of data, reporting, and transparency.*** Platforms took a variety of steps to provide consumers with new features and tools (e.g. Verified Information Features). There have also been several good practices, such as the Trust Project and the Journalism Trust Initiative, that can be more widely implemented and act as a minimum standard for all platforms to live up to. However, overall, there has been insufficient reporting to determine how effective these tools

<sup>484</sup> Plasilova et. al, *Study for the “Assessment of the implementation of the Code of Practice on Disinformation,”* 8.

<sup>485</sup> Twitter expanded their ban of advertising by state media to now include a platform-wide ban on political advertising. It even went as far as introducing an authorization procedure for ‘caused-based advertising’ that seeks to “educate, raise awareness, and/or call for people to take action in connection with civic engagement, economic growth, environmental stewardship, or social equity causes”. Google required verification and ‘paid for by’ disclosure in advance of the European Elections, while Microsoft updated their advertising policies to prohibit political ads globally, including issue-based advertising. Facebook implemented mandatory disclaimers for political and issue-based adverts and went further on the commitment of issue-based advertising than any of the other platforms by listing specific topics that it requires verification for.



were. There has also been a lack of detailed information on the integration of third-party trustworthiness indicators, and a lack of initiative by platforms to seek out cooperation with independent factchecking initiatives. Finally, coherence between platforms remains an issue: for instance, user-friendly developments, such as uniform procedures available on all platforms for users to flag disinformation and common ways of letting users know they have encountered disinformation, have not emerged.

***The impact of the Code was arguably weakest in empowering the research community, with access to data being an ongoing issue for factchecking and research activities – something the Code was designed to address.*** While platforms implemented some policies and tools on this front, most are plagued with concerns about the quality of APIs and datasets made available to researchers. Outsiders also felt that the GDPR was sometimes used by the platforms as an excuse not to engage in information exchange with outside parties. While Vice-President Vera Jourova stated that this should not be the case,<sup>486</sup> a more detailed opinion with guidelines from the European Commission or from the European Data protection Supervisor would be welcomed.<sup>487</sup>

Summarizing the effectiveness of the Code, it can be concluded that it has been successful at bringing together key platforms and increasing the communication between them and stakeholders on an issue as complex as disinformation. However, it is clear to most observers that most of the fundamental pillars of the Code have not (to date) been satisfactorily addressed by the signatories. But, perhaps, the key issue here may not be the rigor of *measures or actions* in the Code, but rather the *regulatory format* in which it was adopted, bringing us to the second question: whether the regulative approach taken by the EU can be considered effective.

A key feature of the EU Code is its self-regulatory nature, which has seen substantial debate surrounding whether it is the right approach to take in this context. External commentators have a number of issues with the self-regulatory approach: Paul-Jasper Dittrich wrote in 2019 that while the self-regulatory approach of the EU Code was effective in coercing industry support, “these measures have not created enough transparency about how the companies are dealing with disinformation and have not led to publicly verifiable results about their success.”<sup>488</sup> The Institute for Strategic Dialogue shared such opinions, stating that “overwhelmingly, the Code of Practice proved the limits of voluntary efforts to bring about systemic change from the signatory companies,”<sup>489</sup> and arguing that a lack of follow-through and enforcement of commitments plagued the otherwise well-documented and formalized Code of Practice.<sup>490</sup> James Pamment, in an early 2020 paper, also notes that the EU Code has

486 European Commission, “Disinformation: EU assesses the Code of Practice and publishes platform reports on coronavirus related disinformation,” September 10, 2020, [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_20\\_1568](https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1568).

487 The May 2021 EC report titled *Guidance on Strengthening the Code of Practice on Disinformation* notes these concerns, and also identifies a path forward. In particular, they note that the European Digital Media Observatory (EDMO) is currently exploring “the possibilities of a code of conduct under Article 40 of the GDPR aimed at ensuring the proper application of privacy and data protection requirements to the sharing of personal data by platforms with researchers.” This would, accordingly, “reduce legal uncertainties and risks for platforms providing access to data and ensure a secure and harmonised environment for processing of personal data for research purposes.” EC, *Guidance on Strengthening the Code of Practice on Disinformation*, 18.

488 Paul-Jasper Dittrich, *Tackling the spread of disinformation: Why a co-regulatory approach is the right way forward for the EU* (Berlin: Hertie School Jacques Delors Centre, 2019), 4.

489 Chloe Colliver, *Cracking the Code: An Evaluation of the EU Code of Practice on Disinformation* (London: Institute for Strategic Dialogue, 2020), 12.

490 Ibid.

produced “mixed results,” noting how few stakeholders appear fully satisfied with the process of self-regulation.<sup>491</sup>

However, more damaging is perhaps the European Commission’s own recognition of the issues with a self-regulatory approach. In their 2020 assessment of the EU Code, while they positively note that the Code’s flexibility can encourage stakeholders to agree and implement the code,<sup>492</sup> they also realize three key limitations of the self-regulation model:

1. **Limited participation:** While the EC did praise the Code’s uptake amongst platforms, they noted “the voluntary nature of the Code establishes an inherent “regulatory asymmetry” as between Code signatories and non-signatories.”<sup>493</sup> The advertising sector was also noted to be largely absent or hesitant about the code, preferring their own initiatives like the Global Alliance for Responsible Media (GARM).<sup>494</sup>
2. **Oversight, monitoring and enforcement:** The EC’s report stated that “the existing self-regulatory framework does not establish an independent oversight mechanism for monitoring the completeness and impact of the signatories’ actions in tackling disinformation.”<sup>495</sup> They also mention that the EU Code lacks any real forms of enforcement mechanisms due to its self-regulatory nature.
3. **Protection of fundamental rights and mechanisms for redress:** Finally, the EC notes that while the Code acknowledges fundamental rights, it “does not set out procedures to ensure in practice the protection of these rights in the pursuit of actions addressing disinformation.”<sup>496</sup>

Ultimately, as both independent experts and the EC themselves have noted, there are a number of limitations to the self-regulation model which should be addressed in future iterations of the Code. The EC recognized this, and there are plans to update the Code of Practice in the near future. For instance, in the December 2020 European Democracy Action Plan, it was stated that “In the Commission’s view, a more robust approach based on clear commitments and subject to appropriate oversight mechanisms is necessary to fight disinformation more effectively.”<sup>497</sup> A document outlining guidance for updating the Code of Practice on Disinformation is expected to be released in Spring 2021.<sup>498</sup> Chapter 6 will build off such observations and present a number of regulatory options available to governments and the EU, ultimately finding that a coregulatory approach would be particularly effective.

## 9.2.4 Recommendations

Based on the analysis of the self-regulatory approach Code of Practice, the following recommendations are made to inform the coregulatory approach for the counter-disinformation regime complex described in Chapter 7.

<sup>491</sup> James Pamment, *EU Code of Practice on Disinformation: Briefing Note for the New European Commission* (Washington, DC: the Carnegie Endowment for International Peace, 2020) [https://carnegieendowment.org/files/Pamment\\_-\\_EU\\_Code\\_of\\_Practice.pdf](https://carnegieendowment.org/files/Pamment_-_EU_Code_of_Practice.pdf).

<sup>492</sup> EC, *Assessment of the Code of Practice on Disinformation*, 17.

<sup>493</sup> Ibid.

<sup>494</sup> Ibid., 18; World Federation of Advertisers, “Global Alliance for Responsible Media,” accessed May 11, 2021, <https://wfanet.org/leadership/garm/about-garm>.

<sup>495</sup> EC, *Assessment of the Code of Practice on Disinformation*, 18.

<sup>496</sup> Ibid., 18-19.

<sup>497</sup> European Commission, *On the European democracy action plan* (Brussels: 2020), 22.

<sup>498</sup> Ibid., 23.

**Focus on transparency.** The majority of the areas where the Code signatories fall short deal with transparency. Platforms have demonstrated that they are reluctant and, occasionally, outright unwilling to provide statistics, datasets, and insights into the effectiveness and operation of their tools and mechanisms to combat disinformation. Moreover, in many cases where platforms are providing data, there remain many concerns with the completeness or depth of that data. The current lack of transparency is especially problematic for the Code, not only because it was a prime focus within several of their pillars, but also because it inhibits the ability of the EC and civil society to monitor the implementation and effectiveness of all the other aims, pillars, and actions of the Code. Finally, as external observers have also noted, more transparency is in the public interest - at its core, it should increase the privacy, safety and rights of the consumer.<sup>499</sup>

**Facilitate actual multistakeholder cooperation.** While the mere existence of a formal feedback mechanism from civil society members should be applauded, the Sounding Board of the Code fell short in terms of adequate representation. While the original intention was to have representatives from the media, civil society, factchecking community and academia, in actuality the Sounding Board severely lacked representation from organizations other than media associations or federations.<sup>500</sup> In a complex domain such as disinformation, the standard-setting process would benefit from the technical, legal, policy and civil rights expertise that is readily available when one broadens the actors involved. It would not only benefit the legitimacy of the outcomes, but would also help build critical mass to advance those standards, make sure that companies adhere to them, and report on possible violations.

**Develop commonly-shared terminologies to encourage platforms to increase collaboration in policy development.** Collaboration starts with developing shared definitions. This is reflected by a key issue realized by the EC: platforms generally are not working together to achieve the goals of the Code, preferring to develop and use their own unique policies, definitions, and approaches. This is problematic: there are many issues in which there is no alignment between the minutiae of the policies, leading to blind spots. Each platform prefers to use their own definition about what constitutes disinformation, including information warfare, influence operations, hybrid warfare, coordinated inauthentic behavior, computational propaganda, and so on. As some have noted, "Without agreement over a definitive EU terminological apparatus for all stakeholders to report against, opaqueness and obfuscation will continue to hamper meaningful progress."<sup>501</sup> Should this common starting point be in place, greater policy alignment and cooperation should also then follow.

**Harmonization of common reporting templates.** The reports delivered to the EC summarizing the efforts by platforms to implement the Code would benefit greatly from more standardization. In their current form, the length and content of the reports vary greatly – for instance, in January 2019, Mozilla's report was only 4 pages, while Google submitted 17 pages.<sup>502</sup> While the EC does request platforms to include their progress on each pillar of the Code within these reports, platforms are free to respond to this in any open-ended way they desire.

<sup>499</sup> Colliver, *Cracking the Code*.

<sup>500</sup> At most of the meetings, only one academic was present. EC, "Meeting of the Multistakeholder Forum on Disinformation."

<sup>501</sup> EC, "Meeting of the Multistakeholder Forum on Disinformation," 2.

<sup>502</sup> Google, *EU Code of Practice on Disinformation: Google Report*, (2019), [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-5/google\\_-\\_ec\\_action\\_plan\\_reporting\\_CF162236-E8FB-725E-C0A3D-2D6CCFE678A\\_56994.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-5/google_-_ec_action_plan_reporting_CF162236-E8FB-725E-C0A3D-2D6CCFE678A_56994.pdf); Mozilla, *Update on Milestones for the Implementation of the Code of Practice on Disinformation* (Belgium: Mozilla, 2019), [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-5/mozilla\\_cop\\_report\\_-\\_18\\_01\\_19\\_CF162508-CF98-8ACD-89BCCC1BA4230DD9\\_56995.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-5/mozilla_cop_report_-_18_01_19_CF162508-CF98-8ACD-89BCCC1BA4230DD9_56995.pdf).

**Develop a set of measurable Key Performance Indicators (KPIs).** What you measure and report on creates incentives. For the next phase of the Code, the EC should therefore develop, in a multistakeholder setting, common and standardized metrics and Key Performance Indicators (KPIs) which each platform should report on under each pillar in their implementation reports. This should go beyond, for instance, KPIs like the absolute number of takedowns, and include measures that assess the quality of a platform's content moderation process. This would help gauge progress and provide evidence towards more comprehensive assessments of the Code's impact as well as improve the differentiated implementation across the pillars, platforms and member states. The 2020 assessment report of the Code of Practice that was commissioned by the EC proposes a cogent and well-argued set of structural indicators (focusing on outcomes) and service-level indicators (focusing on results) to measure the progress of each platform, and could serve as inspiration or as a starting point.<sup>503</sup> KPIs are usually designed and tested against the RACER criteria, i.e. they have to be Relevant, Accepted, Credible, Easy to monitor, and Robust.<sup>504</sup> They should also be (re)evaluated periodically to ensure their compliance with this criteria.

**A noncompliance mechanism.** Even before the Code was brought into power, prominent actors involved in the drafting of it stressed that a noncompliance mechanism could be beneficial, with the Code's own Sounding Board noting the lack of a compliance mechanism among its key concerns about the endeavor in 2018.<sup>505</sup> This foresight proved sound: some of the key areas in which the Code is currently lacking, such as monitoring, oversight, and enforcement, could rationally be dealt with using a noncompliance mechanism which penalizes signatories. This sentiment was shared by the assessment of the Code commissioned by the EC, concluding that "... although the Code of Practice is a self-regulatory instrument – and the first of its kind – introducing a mechanism for action in case of non-compliance of the (insufficient) implementation of the commitments that platforms signed up to, could be considered to enhance the credibility of the agreement. To that effect, the Commission should consider proposals for co-regulation within which appropriate enforcement mechanisms, sanctions and redress mechanisms should be established."<sup>506</sup>

**Move from self-regulation to co-regulation.** Co-regulation is a step further than introducing noncompliance mechanisms. In essence, it would see the Code retaining industry involvement in drafting and creating standards and norms, but would back these up with a legal framework forcing compliance. The next chapter will explore some of the options for a co-regulation model in the context of disinformation. Paul-Jasper Dittrich proposes a simple model of how this might operate, with three main components: a first 'statutory' layer where the EU develops a legislative act, a second 'co-regulatory' layer where industry and stakeholders interact to develop the measures and principles embedded in the legislative layer, and a third layer where companies develop measures for implementation.<sup>507</sup> The EU is no stranger to such approaches: Marsden, Meyer, and Brown identify a number of European co-regulatory schemes.<sup>508</sup> However, they also note that there are issues with co-regulation, primarily that

503 Plasilova et. al, *Study for the "Assessment of the implementation of the Code of Practice on Disinformation,"* 89-95.

504 European Commission, *Indicators to measure Social Protection Performance: Implications for EC Programming* (Brussels: Directorate-General for Development and Cooperation, 2017), 23.

505 The Sounding Board, "The Sounding Board's unanimous final opinion on the so-called code of practice."

506 Plasilova et. al, *Study for the "Assessment of the implementation of the Code of Practice on Disinformation."*

507 Dittrich, *Tackling the spread of disinformation*, 7..

508 Notable internet examples include EURID (which operates .eu domain registries since 2003) and Nominet (another domain registry). Marsden et al, "Platform values and democratic elections." Others, such as Dittrich, also note that the EU approach to countering illegal hate speech online is another good example of a co-regulatory framework. Dittrich, *Tackling the spread of disinformation*, 7.

coregulatory schemes often follow a path to becoming state regulation (and thus lose the industry involvement).<sup>509</sup> Co-regulation does seem to offer a concrete way to pressure platforms to address accountability, transparency, and non-compliance issues that persist in the current self-regulatory approach.

### 9.2.5 The Future of the Code

On May 26<sup>th</sup>, 2021, the European Commission released the *European Commission Guidance on Strengthening the Code of Practice on Disinformation*, a document outlining the EC's recommendations for the future of the Code of Practice on Disinformation.<sup>510</sup> Within this document, they reflected upon their earlier 2020 Assessment of the Code of Practice on Disinformation, and, much like this section, isolates a number of recommendations to address a number of failings within the five key thematic areas of the Code.<sup>511</sup> Like our analysis, the EC takes this prior work from 2020 and accepts all the shortcomings that report finds concerning the Code. As a result, they detail several broad steps the code should take (expanding the Code's scope, broadening the amount of participants, tailoring commitments further for signatories, and the integration of EC initiatives like the European Digital Media Observatory (EDMO) and the Rapid Alert System), as well as a number of detailed steps forward for each individual pillar. Overall, many of the recommendations made by the EC mirror the concerns this case study found, as well as the independently-achieved recommendations made in the preceding section.

So, what will be the impact of the *Guidance on Strengthening the Code*? For starters, the EC does stress that these are but recommendations, albeit ones that they will presumably push hard for platforms to accommodate in a potential revision of the Code. In fact, while the EC is calling for a revision of the Code (largely based on these concerns), there is no guarantee that platforms will even entertain or accept these recommendations.

Notably, one key recommendation proposed by this report that is absent from the EC *Guidance on Strengthening the Code* is the call for the Code to shift from a self-regulatory model to a coregulatory model, which would entitle the EC to much greater powers to actually implement and enforce the recommendations they make. While this is admittedly a large shift, it is one which would perhaps have the greatest impact. Moreover, the EC does recognize the potency of coregulatory regimes: within the *Guidance*, they explicitly identify the DSA as a coregulatory regime, and note that it sets forth a number of new legally obligated measures around transparency, content moderation, and advertising.<sup>512</sup>

All in all, while the EC has made a number of relevant and insightful recommendations which are largely satisfactory in terms of their content, the real question remains how much impact these recommendations will have. After all, the important thing in a self-regulatory regime is not the quality of recommendations presented to it, but rather the signatories willingness to change.

<sup>509</sup> Marsden et al, "Platform values and democratic elections," 16.

<sup>510</sup> EC, *Guidance on Strengthening the Code*.

<sup>511</sup> The *Guidance on Strengthening the Code* also evaluates their COVID-19 monitoring programme, finding that there were five key shortcomings: the quality of reporting, lack of KPIs, lack of independent assessment, the lack of sufficient fact-checking coverage, and the continued monetization of disinformation through advertisement placements. Ibid, 4.

<sup>512</sup> EC, *Guidance on Strengthening the Code*, 2.

# 10 Annex III: The Regulatory Regimes

Disinformation on platforms has seen actors and stakeholders adopt and endorse a variety of regulatory approaches to achieve effects. In this annex, we will discuss these approaches; as well as identifying examples and highlighting both pros and cons of each approach. Finally, we will use this analysis to recommend an ideal path forward for an approach on disinformation.

This section will draw heavily from the work of Chris Marsden, Trisha Meyer, and Ian Brown, who took a more detailed look at regulatory regimes and disinformation in a 2020 article.<sup>513</sup> They identified six main options for regulation on disinformation (see Table 10).

Table 10: Typology of regulation

Option and form of regulation		Typology of regulation
0.	Status quo	Corporate social responsibility, single-company initiatives
1.	Non-audited self-regulation	Industry code of practice, transparency reports, self-reporting
2.	Audited self-regulation	European Code of Practice of September 2018; Global Network Initiative published audit reports
3.	Formal self-regulator	Powers to expel non-performing members, dispute resolution ruling/arbitration on cases
4.	Co-regulation	Industry code approved by Parliament(s) or regulator(s) with statutory powers to supplant
5.	Statutory regulation	Formal regulation – tribunal with judicial review

## 10.1 Option 0 – Status Quo.

*Eg. The United States*

Also commonly referred to as an option of ‘no regulation’, this option entails that there is no formal cooperation between platforms, the private sector, and governments. Instead, platforms and the private sector are left to their own devices to control disinformation, motivated by the libertarian ideals of competition and free markets. The United States is the largest proponent of such an approach, and we can see this manifested in the largely unilateral approaches to disinformation taken by American platforms and the general absence of government intervention, partially a reflection of early interpretation of US law, in particular the famous Section 230 of the US Communication Decency Act.

<sup>513</sup> Marsden et al, “Platform values and democratic elections,” 12.



Experts like Marsden, Meyer, and Brown argue that this option is not one which is suitable for European policymakers (hence why they classified it as an 'option 0'). While environments like the United States focus on the primacy of the right to free speech (hence limiting the influence government can have on civil rights, such as freedom of speech and press), the European approach instead endorses more positive obligations, thus permitting states to intervene in order to protect those rights.<sup>514</sup> In addition to this theoretical explanation, Marsden, Meyer, and Brown also note that, practically speaking, this model is unappealing to Europeans because it fell short during the 2016 American election which saw major issues in the US with bot accounts, disinformation and political advertising. Thus, some form of a regulatory approach on disinformation is ultimately desirable for Europe.

As a sidenote, it should also be noted that platform-initiated oversight mechanisms, such as the Facebook Oversight Board, would also fall under this category. These are mechanisms created by the industry itself, often establishing a higher level of oversight or accountability over their actions. While few and far between, many do consider these good moves by the industry, but critics still wonder about the implications of platforms creating their own quasi-judicial mechanisms rather than using official ones.

## 10.2 Option 1 – Self-regulation

*Eg. EU Code of Practice on Disinformation*

Self-regulation is a key concept in regulation where the industry itself, or in coordination with government and other stakeholders, creates its own standards to which it will hold itself accountable. Key to this is the voluntary nature: there are typically no formal enforcement mechanisms to hold industry accountable to maintaining to the agreed-upon behavior. Instead, the thinking is that since industry collectively decides the obligations, these should be steps they should be willing to make.

Marsden, Meyer and Brown distinguish between three forms of self-regulation: non-audited self-regulation, audited self-regulation, and formal self-regulation (see Table 11).

*Table 11: Types of Self-Regulatory Regimes*

Type of self-regulation	Distinguishing features	Example
<b>Non-audited</b>	Industry-government coordination, but no sanctions or formalized transparency process (other than potentially self-reporting).	Santa Monica Principles on Content Moderation; EU Code of Practice on Disinformation
<b>Audited</b>	Members are subject to regular, independent audits to judge compliance to agreed-upon criteria.	Potentially the Global Network Initiative (GNI), <sup>515</sup> INHOPE.
<b>Formal</b>	Can expel non-complying members; dispute resolution/arbitration on cases; closely supported by existing legislation as the decisive body on issues within its mandate.	Video/Video game ratings (i.e. Pan European Game Information (PEGI), International Age Rating Coalition (IARC))

<sup>514</sup> Ibid., 12.

<sup>515</sup> Marsden et al. state that while the GNI claims to have audited reports, there is little evidence of that on their website and in other publicly available documents. Marsden et al., "Platform values and democratic elections," 14.

Self-regulation has both advocates and critics. Of the former, many note that industry themselves, especially in the rapidly-developing tech industry, are better positioned than the government to develop meaningful and effective legislation, justifying their role in such initiatives.<sup>516</sup> Others note that self-regulatory schemes are effective at bringing together industry partners: this was one of the key benefits identified, for instance, of the self-regulatory nature of the EU Code of Practice on Disinformation.<sup>517</sup> There is also the economic argument: self-regulation is cheaper for both the industry players as well as the government (and, consequently, also the taxpayer and consumer). Finally, self-regulation can be appealing as it helps avoid transnational problems that may arise.<sup>518</sup> However, critics often point out the shortcomings of the free market and that the lack of incentives or enforcement makes it easy for platforms to renege on their commitments to any self-regulatory schemes: something noted by many observers of the EU Code of Practice.<sup>519</sup> Moreover, these critics also note that transparency is simply not sufficiently ensured by voluntary measures alone.<sup>520</sup>

While self-regulation has been criticized for disinformation, a number of relatively successful examples of self-regulation exist in the context of cybersecurity and Internet governance standard-setting by the very industry and civil society organizations that, unlike governments, are in charge of developing or managing the respective products or services. For Internet governance standards, the mantra ‘if it ain’t broke, don’t try to fix it’ still very much applies as it keeps government control over the core Internet protocols and resources at bay. In the cybersecurity context, however, there is a growing concern that the free market is not able to address the many security issues by itself, leading to more government interventions and regulation.

## 10.3 Option 2 – Coregulation

*Eg. Nominet, EURID.*

Whereas in self-regulation industry sets the standards and in regulation government sets the standards, in coregulation this responsibility is shared.<sup>521</sup> To put it more formally, coregulation retains the industry involvement in developing standards, but adds a “statutory underpinning and legitimacy of parliamentary approval for regulatory systems, together with general principles of good regulation” such as audits, an enforcement mechanism, and an appeal process.<sup>522</sup> Many liken it to a pyramid, in which a top layer of regulatory principles set by legislation is interpreted by an independent coregulatory body. This body is made up of a combination of industry, government, and civil society stakeholders, who translate those principles into a regulator design that is implemented through industry-shaped rules and standards, which are monitored by an independent monitoring board.<sup>523</sup>

516 Dennis D. Hirsch, “The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-Regulation?” *Seattle University Law Review* 34, no. 2 (2011): 458.

517 EC, *Assessment of the Code of Practice on Disinformation*.

518 Joseph A. Cannataci & Jeanne Pia Mifsud Bonnici, “Can Self-regulation Satisfy the Transnational Requisite of Successful Internet Regulation?” *International Review of Law, Computers, and Technology* 17 (2003): 53.

519 Colliver, *Cracking the Code*.

520 Dittrich, *Tackling the spread of disinformation*, 4.

521 Hirsch, “The Law and Policy of Online Privacy,” 465.

522 Marsden et al, 14.

523 Dittrich, *Tackling the spread of disinformation*, 9; Marsden et al, 15.

Many advocates of coregulatory approaches argue that they offer many of the same benefits as self-regulatory approaches (such as industry-led implementation and creation of measures) while ensuring greater transparency and compliance due to the present enforcement mechanism. This is especially key for addressing issues like disinformation, where the fast-moving and changing nature of the phenomena demands close collaboration between industry and their regulators. Yet, coregulation does have potential issues. Some note that coregulation is often only a stone's throw away from being turned into full-fledged government regulation – a path often taken by former coregulatory regimes.<sup>524</sup> This can make platforms and industry more hesitant to embrace such schemes. Finally, others still doubt whether a coregulatory approach can have platforms act in a way that is not dominated by business concerns, leading to weaker-than-desired standards.<sup>525</sup>

Yet, these concerns have not deterred many from endorsing co-regulatory approaches to disinformation. Most notably, the EU has openly been exploring how to transition their existing EU Code of Practice on Disinformation from a self-regulatory scheme to a coregulatory one.<sup>526</sup>

## 10.4 Option 3 – Statutory Regulation

*Ex. UK Online Harms (Proposed scheme)*

A traditional regulatory option would see the government assuming full responsibility for creating, drafting, and enforcing regulation. Often, this would entail establishing or appointing a regulatory body with legislative powers to define and enforce acceptable behavior. These efforts can be done either nationally or even regionally.

While traditional regulation is often used and employed by governments to establish and enforce standards and practices in a variety of fields, there are a variety of drawbacks, especially concerning government control over free speech. In specific, the rapidly developing and changing nature of technologies and attack vectors often necessitates the involvement of the actors that are on the frontline and responsible for their implementation in the decision-making process. This need is also crucial when it comes to content moderation, where the platforms are not only the object regulation, but also as active participants and collaborators: something that is difficult if proposed regulation is perceived as unfavorable or undesirable by industry.

Despite these difficulties, some governments have been developing regulatory schemes specifically for disinformation. One such prominent initiative is the emerging Online Harms regime by the United Kingdom, which also hopes to appoint a regulator to implement, oversee and enforce a regulatory framework combatting disinformation that platforms must abide by.<sup>527</sup>

<sup>524</sup> Marsden et al, 15.

<sup>525</sup> Hirsch, "The Law and Policy of Online Privacy," 468.

<sup>526</sup> "We will move from self-regulation to co-regulation," EU Commission vice-president Vera Jourova. Eszter Zalan, "EU Commission plans sanctions on disinformation," *EUobserver*, December 4, 2020, <https://euobserver.com/political/150279>.

<sup>527</sup> Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department, *Online Harms White Paper* (UK: HM Government, 2019), 53.

# 11 Annex IV: ISACs - the Gold Standard in Cybersecurity Information Exchange

Information Sharing and Analysis Centers (ISACs), sometime known as WARPs, have formed a backbone of cybersecurity for nearly 20 years. First developed in the United States as a linchpin in developing critical infrastructure protection (created by Presidential Decision Directive 63 (PDD-63) in 1998 and updated in 2003 with Homeland Security Presidential Directive/HSPD-7), ISACs represent the focus for public-private-partner-partnerships in cybersecurity in virtually all liberal democracies. ISACs are nonprofit organizations that function as entrusted entities that facilitate two-way information exchange on threats and vulnerabilities between private and private actors (and civil society when appropriate) and provide members with analysis, early warnings, and tools to mitigate risks, enhance mutual trust and resilience, and improve situational awareness. Underpinning these activities in making sure all members work from a similar standard, are capability building and awareness raising at all levels – e.g. through training. Overall, ISACs can be country-focused, sector-specific, or follow an international collaborative model. In the US, most ISACs are clustered around sectors that are provided and staffed by industry partners. They vary greatly in size and efficiency. This chapter explains the core components of an ISAC, including its overall functions, the actors and overall roles, possible organizational structures, and the most common funding sources.

## 1. Functions

Overall, an ISAC provides three core functions: information sharing lies at the heart, followed by capacity and trust building.

### 1.1. Information sharing

Most of the information exchange focuses is about threats, vulnerabilities, mitigation, situational awareness, and best practices or tools, that is shared through special web-based platforms or (encrypted) email lists using a set template and the Traffic Light Protocol (TLP), or during meetings. In most cases, the information is usually validated before it is widely shared, potentially anonymizing the original sender, while in other cases it can be shared directly by the sender through an email list. Some ISACs have an inner circle of the leading members in which information is shared in more detail. Often done through a formalized agreement that

specify the means and the type of information shared. Special regulation for public administration might inhibit them from signing such an agreement. One of the main challenges for smaller members is they lack the resources to analyze all the information shared by an ISAC. The ISAC secretariat, a governmental body or vetted civil society experts can analyze the data can alleviate this concern.

### 1.2. Capacity building

Capacity building allows members to increase the overall expertise and security of the community as a whole and contribute to the level and quality of participation of smaller members. While it can be applied to nearly every facet of an organization, it usually focuses on providing threat analysis, training and exercises, and technical tools. Threat analysis streamlines the vast amount of threat information that smaller members need to consume. Training and exercises raise the overall awareness, knowledge and expertise and of the members, and contribute to trust building. Finally, sharing technical tools in a centralized setting can raise the technical capacity of new members.

### 1.3. Trust building

Trust lies at the core of an ISAC. The more often members interact, the higher the level of trust and the overall collaboration and cooperation. Trust will mainly be tested whether members actually share credible information and act on that information in a timely fashion. It can be facilitated in several ways. This includes technical tools, such as the TLP protocol, and formal tools, as a terms of reference, non-disclosure agreements, or code of conduct can contribute to trust. The formal agreements can set out the obligations of the members when it comes to sharing and handling information and establish penalties for misconduct.

## 2. Actors/Membership

ISACs are often constituted of stakeholders from government and industry, and often involve civil society:

1. **Governments'** role is often twofold: first, facilitating the ISAC through hosting or funding, and second, creating a legal framework or mandate for the establishment of the ISAC and information sharing (not limited to legal regulation, but can include standards, government programs and strategies). Public administration might not just be involved in a secretarial function, but its defensive operational branch can also actively participate in the information sharing and analysis functions of the ISAC, much like the NCSC in the Netherlands that furthermore functions as a focal point for incident reporting and handling for critical sectors. The law enforcement and intelligence community, on the other hand, is not directly involved due to their classified nature that could jeopardize information sharing. They can, however, have an indirect link as a partner in dedicated sessions. In the traditional ISAC model, public administration often has a different role than industry mainly having to do with the sharing of classified information, but in the end both stakeholders should adhere to same set of rules.
2. **Industry** is the primary driving force of an ISAC because of their position at the front line of defense, own most of the infrastructure and products, because a higher level of cybersecurity is essential to its business interests and continuity, and finally because of legal

obligations to report incidents and ensure continuity of critical services. Industry is therefore often a facilitator or a member that, more so than public administration, determines the *modus operandi* for cooperation in an ISAC and its level of engagement.

- 3. Civil society:** Civil society members, in particular academia, technical experts or news agencies, can be involved in some ISACs as a partner. For academics, it offers a two-way information exchange with industry. The latter can communicate research needs, while academics and technical experts provide expertise and possibilities for new solutions that can be useful for the industry members. This anchors the research in the technical reality of the social media companies.

As a general rule, there are often three different roles within an ISAC:

- 1. The facilitator** functions as the secretariat of the group and takes on operational support and logistics of the group, in a similar way as the GIFCT Secretariat does so. This can be done by a government agency or a company.
- 2. The members** often consist of industry parties that actively share and receive information and possibly pay a membership fee.
- 3. The partners** can participate in dedicated sessions usually to offer specific information (e.g. researchers, law enforcement or intelligence community providing cybersecurity expertise or intelligence) or to discuss a specific topic (how the standards proposed can be transposed or should be interpreted).

### 3. Organizational structures

ISACs can be governed in many ways depending on the needs and availability of the stakeholders, as well as the overall objective and the tools that support its core functions. Sometimes they can have clear structures and well-defined roles with a dedicated secretariat or management board, while in other cases it lacks a formal structure and focuses on ad-hoc voluntary information exchange.

- 1. The structured approach** includes a clear organizational structure that includes a management board, steering committee, or a chair. Those roles are rarely elected and limited to a tenure (1-2 years). The rule of thumb is that the leading positions are often assigned to the industry actors most involved. In the case of disinformation, this would mean large social media platforms, such as Facebook, Twitter, or YouTube. Afterwards, a strategy and action plan is set up to establish the goals, direction, structures, terms of reference, and election rules of the community. As a potential add-on or separate model altogether, a supporting body in the form of a secretariat can be added as a facilitator. This function is often taken up by the public sector (if it's involved). According a study of ENISA that surveys a wide range of ISAC members, especially international ISACs benefit from a secretariat role to manage the operational support for and communication.
- 2. The flexible approach** has no clearly-defined roles, structure, action plan, and is often governed or managed by members on a voluntary basis in which decisions are made on an ad-hoc basis. It enables members within a community to familiarize itself with the



organization culture and challenges of other members, but the lack of formality can lead to low stakeholder engagement.

## 4. Funding

There are several ways to fund an ISAC. The most common sources are mandatory fees (can motivate stakeholders to more actively contribute), voluntary contributions (both cash and in kind), and government subsidies (cash contribution to stimulate industry funding and cooperation, or in kind support, such as a secretariat function).

# 12 Annex V: Interviewees

Interviews have been conducted with stakeholders from civil society, government, and industry, to receive feedback on the proposals of this report, for which we would like to express our thanks:

- Representatives Ministry of Foreign Affairs, Netherlands
- Representatives Ministry of Defense, Netherlands
- Representatives Ministry of Economic Affairs and Climate, Netherlands
- Representative Ministry of Interior Affairs, Netherlands
- Representative Facebook
- Representative Google
- Representative Microsoft
- Representative European External Action Service
- Representative German Federal Foreign Office
- EU Officials
- Bart Groothuis, Member of European Parliament
- Anne van Heijst, Policy Advisor at European Parliament
- Arthur de Liedekerke, Ministry of Defense, France

# Bibliography

- @Delbius. "Help us shape our approach to synthetic and manipulated media." *Twitter*, November 11, 2019. [https://blog.twitter.com/en\\_us/topics/company/2019/synthetic\\_manipulated\\_media\\_policy\\_feedback.html](https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html).
- @policy. "Update on the Global Internet Forum to Counter Terrorism." *Twitter*, December 4, 2017. [https://blog.twitter.com/en\\_us/topics/events/2017/GIFCTupdate.html](https://blog.twitter.com/en_us/topics/events/2017/GIFCTupdate.html).
- Africa Check, Chequedo, and Full Fact. "Fact checking doesn't work (the way you think it does)." *Full Fact*, June 20, 2019. <https://fullfact.org/blog/2019/jun/how-fact-checking-works/>.
- "A Guide to the UN GGE". Narrated by James Lewis and Chris Painter. Inside Cyber Diplomacy. *Center for Strategic and International Studies*, June 11, 2021. <https://www.csis.org/node/61229>
- Al-Rawi, Ahmed and Shukla, Vishal. "Bots as Active News Promoters: A Digital Analysis of COVID-19 Tweets." *Information* 11 (2020): 461-474. DOI:10.3390/info1100461.
- Arun, Chinmayi. "AI and the Global South: Designing for Other Worlds." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford: Oxford University Press, 2020. <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397>.
- Attorney General's Office and Wright, Jeremy. "Cyber and International Law in the 21st Century", *Government of the United Kingdom*, May 23, 2018. <https://www.gov.uk/government/speeches/cyber-and-international-law-in-the-21st-century>.
- de Baillencourt, Julie. "Meet TikTok's European Safety Advisory Council." *TikTok*, April 21, 2021. <https://newsroom.tiktok.com/en-eu/meet-tiktok-european-safety-advisory-council-eu>.
- BBC. "Covid misinformation on Facebook is killing people – Biden." *BBC*. July 17, 2021. <https://www.bbc.co.uk/news/world-us-canada-57870778>
- BBC. "Online Safety Bill 'catastrophic for free speech'" *BBC*. June 23, 2021. <https://www.bbc.com/news/technology-57569336>
- BBC. "Twitter to ban all political advertising." *BBC*, October 31, 2019. <https://www.bbc.com/news/world-us-canada-50243306>.
- Beinart, Peter. "The U.S. Needs to Face Up to Its Long History of Election Meddling." *The Atlantic*, July 22, 2018. <https://www.theatlantic.com/ideas/archive/2018/07/the-us-has-a-long-history-of-election-meddling/565538/>.
- Bettadapur, Arjun Narayan. "TikTok partners with fact-checking experts to combat misinformation." *TikTok*, October 1, 2020. <https://newsroom.tiktok.com/en-au/tiktok-partners-with-fact-checking-experts-to-combat-misinformation>.
- Bidar, Musadiq. "Twitter will label posts with misleading information about COVID-19 vaccines." *CBS News*, March 2, 2021. <https://www.cbsnews.com/news/twitter-covid-19-vaccine-misinformation-labels/>.
- Bodle, Robert. "The Ethics of Online Anonymity or Zuckerberg vs "Moot"." *ACM SIGCAS Computers and Society* 43, no. 1 (May 2013). <https://doi.org/10.1145/2505414.2505417>.
- Bontridder, Noemi, and Poulet, Yves. "The role of Artificial Intelligence in disinformation". *Namur Digital Institute, Faculty of Law* (2021): 1-24. <https://researchportal.unamur.be/en/publications/the-role-of-artificial-intelligence-in-disinformation>
- Botcheva, Kalina, Posetti, Julie, Teyssou, Denis, Meyer, Trisha, Gregory, Sam, Hanot Clara and Maynar, Diana. *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet'*. International Telecommunication Union and the United Nations Education, Scientific and Cultural Organization, 2020. [https://www.broadbandcommission.org/Documents/working-groups/FoE\\_Disinfo\\_Report.pdf](https://www.broadbandcommission.org/Documents/working-groups/FoE_Disinfo_Report.pdf)

- Broeders, Dennis, Cristiano, Fabio and Weggemans, Daan. "Too Close for Comfort: Cyber Terrorism and Information Security across National Policies and International Diplomacy". *Studies in Conflict & Terrorism* (2021), DOI: 10.1080/1057610X.2021.1928887.
- Broeders, Dennis. "Creating Consequences for Election Interference". *Directions. Cyber Digital Europe*, May 15, 2020. <https://directionsblog.eu/creating-consequences-for-election-interference>
- Broeders, Dennis. "The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: a mid-process assessment". *Journal of Cyber Policy* (2021). DOI: 10.1080/23738871.2021.1916976
- Budnitsky, Stanislav. "Russia's great power imaginary and pursuit of digital multipolarity." *Internet Policy Review* 9, no. 3 (August 2020). DOI: 10.14763/2020.3.1492.
- Burt, Tom. "New Steps to Combat Disinformation." *Microsoft*, September 1, 2020. <https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/>.
- Cannataci, Joseph A. and Bonnici, Jeanne Pia Mifsud. "Can Self-regulation Satisfy the Transnational Requisite of Successful Internet Regulation?" *International Review of Law, Computers, and Technology* 17 (2003): 51-61. <https://doi.org/10.1080/1360086032000063110>.
- Carden, Meredith. "Responding to The Guardian: A Fact-Check on Fact-Checking." *Facebook*, December 13, 2018. <https://about.fb.com/news/2018/12/guardian-fact-check/>.
- Carr, Madeleine. "Public-private partnerships in national cyber-security strategies". *International Affairs* 92, no. 1 (January 2016). <https://doi.org/10.1111/1468-2346.12504>
- Caswell, Brian; Beale, Jay; and Baker, Andrew. *Snort Intrusion Detection and Prevention Toolkit*. Elsevier Inc, 2006. <https://doi.org/10.1016/B978-1-59749-099-3.X5000-9>.
- Centre for Climate Change Communication. *The Debunking Handbook*. Washington: The George Mason University, 2020. <https://www.climatechangecommunication.org/debunking-handbook-2020/>.
- Centre for Countering Digital Hate (CCDH). *Failure to Act: How Tech Giants Continue to Defy Calls to Rein in Vaccine Misinformation*. CCDH: 2020. [https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9\\_dbc700e9063b4653a7d27f4497f3c2c2.pdf](https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_dbc700e9063b4653a7d27f4497f3c2c2.pdf).
- Chan, Man-pui S.; Jones, Christopher R.; Jamieson, Kathleen H.; and Albarracín, Dolores. "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation." *Psychological Science* 28, no. 11 (2017): 1531-1546. <https://doi.org/10.1177/0956797617714579>.
- Choo, Foo Yun and Paul, Katie. "Facebook launches climate science info center amid fake news criticism." *Reuters*, September 15, 2020. <https://www.reuters.com/article/facebook-climatechange-int-idUSKBN2660M5>.
- Christchurch Call. *The Christchurch Call to Action To Eliminate Terrorist and Violent Extremist Content Online*. <https://www.christchurchcall.com/christchurch-call.pdf>.
- Clegg, Nick. "Combating COVID-19 Misinformation Across Our Apps." *Facebook*, March 25, 2020. <https://about.fb.com/news/2020/03/combating-covid-19-misinformation/>.
- Clegg, Nick. "Facebook's Response to the Oversight Board's First Set of Recommendations." *Facebook*, February 25, 2021. <https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations/>.
- Clegg, Nick. "Welcoming the Oversight Board." *Facebook*, May 6, 2020. <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/>.
- Cloudflare. "What is a CDN? | How do CDNs work?" Accessed May 6, 2021. <https://www.cloudflare.com/learning/cdn/what-is-a-cdn/>.
- Cohen, Harris. "Bringing fact check information to Google Images." *Facebook*, June 22, 2020. <https://www.blog.google/products/search/bringing-fact-check-information-google-images>.
- Cohen, Robin; Moffatt, Karyn; Ghenai, Amira; Yang, Andy; Corwin, Margaret; Lin, Gary; Zhao, Raymond; Ji, Yipeng; Parmentier, P'ng, Jason; Tan, Wil; and Gray, Lachlin. "Addressing Misinformation in Online Social Networks: Diverse Platforms and the Potential of Multiagent Trust Modeling." *Information* 11 (2020). <https://doi.org/10.3390/info11110539>.

- Coleman, Keith. "Introducing Birdwatch, a community-based approach to misinformation." *Twitter*, January 25, 2021. [https://blog.twitter.com/en\\_us/topics/product/2021/introducing-bird-watch-a-community-based-approach-to-misinformation.html](https://blog.twitter.com/en_us/topics/product/2021/introducing-bird-watch-a-community-based-approach-to-misinformation.html).
- Collins, Ben. "Twitter is testing new ways to fight misinformation — including a community-based points system." *NBC*, February 20, 2020. <https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931>.
- Colliver, Chloe. *Cracking the Code: An Evaluation of the EU Code of Practice on Disinformation*. London: Institute for Strategic Dialogue, 2020. [https://www.isdglobal.org/wp-content/uploads/2020/06/isd\\_Cracking-the-Code.pdf](https://www.isdglobal.org/wp-content/uploads/2020/06/isd_Cracking-the-Code.pdf).
- Commission Nationale de Contrôle de la Campagne électorale en vue de l'Élection Présidentielle, "Recommandation aux médias suite à l'attaque informatique dont a été victime l'équipe de campagne de M. Macron", May 6, 2017, <http://www.cncep.fr/communiqués/cp14.html>.
- Condliffe, Jamie. "Snapchat Has a Plan to Fight Fake News: Ripping the 'Social' from the 'Media'." *MIT Technology Review*, November 29, 2017. <https://www.technologyreview.com/2017/11/29/147413/snapchat-has-a-plan-to-fight-fake-news-ripping-the-social-from-the-media/>.
- Conger, Kate. "Twitter Had Been Drawing a Line for Months When Trump Crossed It." *The New York Times*, May 30, 2020. <https://www.nytimes.com/2020/05/30/technology/twitter-trump-dorsey.html>.
- Conger, Kate. "Twitter says it labeled 0.2% of all election-related tweets as disputed." *The New York Times*, November 12, 2020. <https://www.nytimes.com/2020/11/12/technology/twitter-says-it-labeled-0-2-of-all-election-related-tweets-as-disputed.html>.
- Conger, Kate. "Twitter to Label Abusive Tweets From Political Leaders." *The New York Times*, June 27, 2019. <https://www.nytimes.com/2019/06/27/technology/twitter-politicians-labels-abuse.html>.
- Conseil Supérieur de l'Audiovisuel (CSA). *Combating the dissemination of false information on online platforms: an evaluation of the application and effectiveness of the measures implemented by operators in 2019*. Paris: CSA, 2020.
- Cooper, Paige. "How the Facebook Algorithm Works in 2021 and How to Make it Work for You." *Hootsuite*, February 10, 2021. <https://blog.hootsuite.com/facebook-algorithm/>.
- Corn, Gary. "Coronavirus Disinformation and the Need for States to Shore Up International Law." *Lawfare*, April 2, 2020. <https://www.lawfareblog.com/coronavirus-disinformation-and-need-states-shore-international-law>.
- Corn, Gary and Jensen, Eric. "The Technicolor Zone of Cyberspace – Part I." *Just Security* May 30, 2018. <https://www.justsecurity.org/57217/technicolor-zone-cyberspace-part/>.
- Cornell, Joe. "How to Report Videos, Accounts, and Comments on TikTok." *How-To Geek*, February 24, 2020. <https://www.howtogeek.com/658518/how-to-report-videos-accounts-and-comments-on-tiktok/#:~:text=Tap%20the%20three%2Ddot%20icon,a%20description%20of%20your%20report>.
- Council of Europe. *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*. Strasbourg: Council of Europe, 2018. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>.
- Coyne, Bridget. "Helping identify 2020 US election candidates on Twitter." *Twitter*, December 12, 2019. [https://blog.twitter.com/en\\_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html](https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html).
- Coyne, Bridget. "Introducing US Election Labels for Midterm Candidates." *Twitter*, May 23, 2018. [https://blog.twitter.com/en\\_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html](https://blog.twitter.com/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html).
- Credibility Coalition. "American Press Institute Fact-Checking and Accountability Journalism Project." Accessed on May 12, 2021. <https://credibilitycoalition.org/credcatalog/project/american-press-institute-fact-checking-and-accountability-journalism-project/>.
- Cresci, Stefano. "A Decade of Social Bot Detection." *Communications of the ACM* 63, no. 10 (September 2020): 72-83. <https://doi.org/10.1145/3409116>.

- Cruz, Taylor and Simoes, Paulo. *EECWS 2019 18th European Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited: 2019.
- Culliford, Elizabeth. "Facebook to label all posts about COVID-19 vaccines." *Reuters*, March 15, 2021. <https://www.reuters.com/article/us-health-coronavirus-facebook-idUSKBN2B70NJ>.
- Culliford, Elizabeth. "Twitter opens up data for researchers to study COVID-19 tweets." *Reuters*, April 29, 2020. <https://www.reuters.com/article/us-health-coronavirus-twitter-data/twitter-opens-up-data-for-researchers-to-study-covid-19-tweets-idINKBN22B2Q1>.
- Culliford, Elizabeth and Paul, Katie. "With fact-checks, Twitter takes on a new kind of task." *Reuters*, May 31, 2020. <https://www.reuters.com/article/us-twitter-factcheck-idUSKBN2360U0>.
- Cybersecurity and Infrastructure Security Agency. "Foreign Interference." Accessed May 6, 2021. <https://www.cisa.gov/publication/foreign-interference>.
- Darcy, Oliver. "How Twitter's algorithm is amplifying extreme political rhetoric." *CNN Business*, March 22, 2019. <https://edition.cnn.com/2019/03/22/tech/twitter-algorithm-political-rhetoric/index.html>.
- The Digital, Culture, Media and Sport Committee. *Disinformation and 'fake news': Final Report*. London: Parliamentary Copyright House of Commons, 2019. <https://publications.parliament.uk/pa/cm201719/cmselect/cmcmds/1791/1791.pdf>.
- Digital Industry Group Inc. *Australian Code of Practice on Disinformation and Misinformation*. DIGI: 2021. <https://digi.org.au/wp-content/uploads/2021/02/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-PDF-Feb-22-2021.pdf>.
- Directorate-General for Communications Networks, Content and Technology (European Commission); Open Evidence; and RAND Europe. *Study on media literacy and online empowerment issues raised by algorithm-driven media services*. Luxembourg: Publications Office of the European Union, 2019. <https://op.europa.eu/en/publication-detail/-/publication/a9101f97-f940-11e9-8c1f-01aa75ed71a1/language-en>.
- Dittrich, Paul-Jasper. *Tackling the spread of disinformation: Why a co-regulatory approach is the right way forward for the EU*. Berlin: Hertie School Jacques Delors Centre, 2019. <http://aei.pitt.edu/102500/1/2019.dec.pdf>.
- Dizikes, Peter. "The catch to putting warning labels on fake news." *MIT News*, March 2, 2020. <https://news.mit.edu/2020/warning-labels-fake-news-trustworthy-0303>.
- DNS Abuse Framework. "Framework to Address Abuse." May 29, 2020. [https://dnsabuseframework.org/media/files/2020-05-29\\_DNSAbuseFramework.pdf](https://dnsabuseframework.org/media/files/2020-05-29_DNSAbuseFramework.pdf).
- Donovan, Joan. *Navigating the Tech Stack: When, Where, and How Should We Moderate Content?* Waterloo: Centre for International Governance Innovation, 2019. [https://www.cigionline.org/sites/default/files/documents/Platform-gov-WEB\\_VERSION.pdf](https://www.cigionline.org/sites/default/files/documents/Platform-gov-WEB_VERSION.pdf).
- Dorsey, Jack (@jack). "We've made the decision to stop all political advertising on Twitter globally..." *Twitter*, October 30, 2019. [https://twitter.com/jack/status/1189634360472829952?ref\\_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1189634360472829952%7Ctwgr%5E%7Ctwcon%5Es1\\_&ref\\_url=https%3A%2F%2Fwww.bbc.com%2Fnews%2Fworld-us-canada-50243306](https://twitter.com/jack/status/1189634360472829952?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1189634360472829952%7Ctwgr%5E%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.bbc.com%2Fnews%2Fworld-us-canada-50243306).
- Douek, Evelyn. "The Facebook Oversight Board's First Decisions: Ambitious, and Perhaps Impractical." *Lawfare*, January 28, 2021. <https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical>.
- Douek, Evelyn. "The Oversight Board Moment You Should've Been Waiting For: Facebook Responds to the First Set of Decisions." *Lawfare*, February 26, 2021. <https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-facebook-responds-first-set-decisions>.
- Douek, Evelyn. "The Rise of Content Cartels." *Knight First Amendment Institute at Columbia University*, February 11, 2020. <https://knightcolumbia.org/content/the-rise-of-content-cartels>.
- Dreydus, Emily and Lapowsky, Issie. "Facebook Is Changing News Feed (Again) to Stop Fake News." *Wired*, April 10, 2019. <https://www.wired.com/story/facebook-click-gap-news-feed-changes/>.



- Le Drian, Jean-Yves. Interviewed in Le Journal du Dimanche. "France Thwarts 24,000 Cyber-Attacks Against Defence Targets." BBC. January 8, 2017. <https://www.bbc.com/news/world-europe-38546415>.
- Drutman, Lee. "Fact-Checking Misinformation Can Work. But It Might Not Be Enough." *FiveThirtyEight*, June 3, 2020. <https://fivethirtyeight.com/features/why-twitthers-fact-check-of-trump-might-not-be-enough-to-combat-misinformation/>.
- Dunčikaitė, Ieva, Žemgulytė, Deimantė and Valladares, Jorge. *Paying for Views: Solving Transparency and Accountability Risks in Online Political Advertising*. Transparency International, 2021. [https://images.transparencycdn.org/images/2021\\_Report\\_PayingForViews-OnlinePoliticalAdvertising\\_English.pdf](https://images.transparencycdn.org/images/2021_Report_PayingForViews-OnlinePoliticalAdvertising_English.pdf)
- Eddy, Melissa and Scott, Mark. "Facebook and Twitter Could Face Fines in Germany Over Hate Speech Posts." *New York Times*, March 14, 2017. <https://www.nytimes.com/2017/03/14/technology/germany-hate-speech-facebook-tech.html>.
- The Editorial Board. "Fact-Checking Facebook's Fact Checkers." *Wall Street Journal*, March 5, 2021. <https://www.wsj.com/articles/fact-checking-facebooks-fact-checkers-11614987375>.
- EDRI. "RIP CleanIT." January 29, 2013. <https://edri.org/our-work/rip-cleanit/>.
- Elfrink, Tim. "A cesspool of hate: U.S. web firm drops 8chan after El Paso shooting." *The Washington Post*, August 5, 2019. <https://www.washingtonpost.com/nation/2019/08/05/chan-dropped-cloudflare-el-paso-shooting-manifesto/>.
- The Embassy of the Russian Federation to the United Kingdom of Great Britain and Northern Ireland. "Concept of a Convention on International Information Security." November 28, 2011. <https://rusemb.org.uk/policycontact/52>.
- Emery, Chelsea. "Comcast, NetZero agree to block Internet child porn." *Reuters*, July 29, 2008. <https://www.reuters.com/article/us-comcast-childporn/comcast-netzero-agree-to-block-internet-child-porn-idUSN2935028520080729>.
- No Author. "TikTok Introduces Warning Label To Combat Fake News." *Entrepreneur Europe*, February 4, 2021. <https://www.entrepreneur.com/article/364767>.
- Europa Nu. "The Code of conduct on countering illegal hate speech online." June 22, 2020. [https://www.europa-nu.nl/id/vl9qfzaji8mu/nieuws/the\\_code\\_of\\_conduct\\_on\\_countering?ctx=vg9pj7ufwbwe&tab=0](https://www.europa-nu.nl/id/vl9qfzaji8mu/nieuws/the_code_of_conduct_on_countering?ctx=vg9pj7ufwbwe&tab=0).
- European Commission. "Annual self-assessment reports of signatories to the Code of Practice on Disinformation 2019." Last updated March 8, 2021. <https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019>.
- European Commission. *Assessment of the Code of Practice on Disinformation: Achievements and areas for further improvement*. Brussels: European Commission, 2020. <https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement>.
- European Commission. "A draft code of practice on online disinformation." Last updated: March 8, 2021. <https://digital-strategy.ec.europa.eu/en/library/draft-code-practice-online-disinformation>.
- European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Fostering a European approach to Artificial Intelligence*. Brussels: European Commission, April 21, 2021
- European Commission. "Disinformation: EU assesses the Code of Practice and publishes platform reports on coronavirus related disinformation." September 10, 2020. [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_20\\_1568](https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1568).
- European Commission. "The EU Code of conduct on countering illegal hate speech online." June 30, 2016. [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en).
- European Commission. *EU Code of Practice on Disinformation*. September 26, 2018. <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation>.

- European Commission. *European Commission Guidance on Strengthening the Code of Practice on Disinformation*. Brussels: European Commission. May 26, 2021. <https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation>.
- European Commission. "European Democracy Action Plan: making EU democracies stronger". Press release, December 03, 2021. [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_20\\_2250](https://ec.europa.eu/commission/presscorner/detail/en/IP_20_2250).
- European Commission. "FAQ — Digital Services Act." Last updated: December 15, 2020. <https://ec.europa.eu/digital-single-market/en/faq/faq-digital-services-act>.
- European Commission. *Indicators to measure Social Protection Performance: Implications for EC Programming*. Brussels: Directorate-General for Development and Cooperation, 2017.
- European Commission. "Meeting of the Multistakeholder Forum on Disinformation." 2018. <https://digital-strategy.ec.europa.eu/en/library/meeting-multistakeholder-forum-disinformation>.
- European Commission. *A multi-dimensional approach to disinformation Report of the independent High level Group on fake news and online disinformation*. Luxembourg: Publications Office of the European Union, 2018.
- European Commission. *On the European democracy action plan*. Brussels: 2020. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0790&from=EN>.
- European Commission. "Political advertising – improving transparency". *European Commission*. [https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Political-advertising-improving-transparency\\_en](https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Political-advertising-improving-transparency_en)
- European Commission. Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Brussels: European Commission, December 15, 2020. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en>.
- European Commission. "Roadmaps to implement the Code of Practice on disinformation." Last updated: March 8, 2021. <https://digital-strategy.ec.europa.eu/en/news/roadmaps-implement-code-practice-disinformation>.
- European Commission. *Tackling online disinformation: a European Approach*. Brussels: European Commission, 2018. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236>.
- European Court of Auditors. *Disinformation affecting the EU: tackled but not tamed*. Luxembourg: European Court of Auditors, 2021. [https://www.eca.europa.eu/Lists/ECADocuments/SR21\\_09/SR\\_Disinformation\\_EN.pdf](https://www.eca.europa.eu/Lists/ECADocuments/SR21_09/SR_Disinformation_EN.pdf)
- European Regulators Group for Audiovisual Media Services. *ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice*. ERGA: 2020. <https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf>.
- Faesen, Louk; Torossian, Bianca; Mayhew, Elliot; and Zensus, Carl. *Conflict in Cyberspace: Parsing the Threats and the State of International Order in Cyberspace*. The Hague: The Hague Centre for Strategic Studies, 2019. <https://hcss.nl/report/conflict-in-cyberspace-parsing-the-threats-and-the-state-of-international-order-in-cyberspace/>.
- Faesen, Louk; Sweijts, Tim; Klimburg, Alexander; MacNamara, Conor; and Mazarr, Michael. *From Blurred Lines to Red Lines: How Countermeasures and Norms Shape Hybrid Conflict*. The Hague: The Hague Centre for Strategic Studies. September 2020.
- Facebook. "Community Standards Enforcement Report." Last modified February 2021. <https://transparency.facebook.com/community-standards-enforcement#fake-accounts>.
- Facebook. "Coordinated Inauthentic Behavior." Accessed May 12, 2021. <https://about.fb.com/news/tag/coordinated-inauthentic-behavior/>.
- Facebook. "Coronavirus (COVID-19) Informatiecentrum." Accessed on May 12, 2021. [https://www.facebook.com/coronavirus\\_info/](https://www.facebook.com/coronavirus_info/).
- Facebook. *Facebook Baseline Report on Implementation of the Code of Practice on Disinformation*. Facebook: 2019. [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-5/facebook\\_baseline\\_report\\_on\\_implementation\\_of\\_the\\_code\\_of\\_practice\\_on\\_disinformation\\_CF161D11-9A54-3E27-65D58168CAC40050\\_56991.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-5/facebook_baseline_report_on_implementation_of_the_code_of_practice_on_disinformation_CF161D11-9A54-3E27-65D58168CAC40050_56991.pdf).

- Facebook. "Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism." June 26, 2017. <https://about.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/>.
- Facebook. *Facebook report on the implementation of the Code of Practice for Disinformation*. Facebook: 2019.
- Facebook. *Facebook response to the European Commission Communication on Covid-19 Disinformation: Report for December 2020*. Facebook: 2021.
- Facebook. "How Our Fact-Checking Program Works." *Facebook Journalism Project*, August 11, 2020. <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/how-it-works>.
- Facebook. "Inauthentic Behaviour." *Community Standards*. Accessed May 6, 2021. [https://www.facebook.com/communitystandards/inauthentic\\_behavior](https://www.facebook.com/communitystandards/inauthentic_behavior).
- Facebook. "Facebook key milestones for the Implementation of the Code of Practice on Disinformation."
- Facebook. "Launching Our Voting Information Center on Facebook and Instagram." August 13, 2020. <https://www.facebook.com/business/news/launching-our-voting-information-center-on-facebook-and-instagram>.
- Facebook. *Oversight Board Charter*. Facebook: 2019. [https://about.fb.com/wp-content/uploads/2019/09/oversight\\_board\\_charter.pdf](https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf).
- Facebook. "Partnering with Third-Party Fact-Checkers." *Facebook Journalism Project*, March 23, 2020. <https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners>.
- No Author. "Facebook using machine learning to fight fake news." *Internet of Business*, accessed May 12, 2021. <https://internetofbusiness.com/facebook-machine-learning-fake-news/>.
- Facebook. "Working to Stop Misinformation and False News." Accessed May 11, 2021. <https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news>.
- Facebook Business Help Center. "Fact-Checking on Facebook." *Facebook*, accessed on May 12, 2021. <https://www.facebook.com/business/help/2593586717571940?id=673052479947730>.
- Fiedler, Kirsten. "EU Internet Forum against terrorist content and hate speech online: Document pool." *European Digital Rights*, March 10, 2016. <https://edri.org/eu-internet-forum-document-pool/>.
- Fowler, Geoffrey A. and Alcantara, Chris. "Gatekeepers: These tech firms control what's allowed online." *The Washington Post*, March 24, 2021. <https://www.washingtonpost.com/technology/2021/03/24/online-moderation-tech-stack/>.
- France. "Against information manipulation." November 20, 2018. <https://www.gouvernement.fr/en/against-information-manipulation>.
- François, Camille. "Actors, Behaviors, Content: A Disinformation ABC". *Transatlantic Working Group*, September 20, 2019. [https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC\\_Framework\\_2019\\_Sept\\_2019.pdf](https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf)
- French Secretary of State for Digital Affairs. *Creating a French framework to make social media platforms more accountable: Acting in France with a European vision*. Paris: 2019. <https://thecre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf>.
- Finnemore, Martha and Sikkink, Kathryn. "International Norm Dynamics and Political Change." *International Organizations* 52, no. 4 (1998): 887-917. <https://www.jstor.org/stable/2601361?seq=1>.
- Funke, Daniel and Flamini, Daniela. "A guide to anti-misinformation actions around the world." Poynter. Last updated: August 13, 2019. <https://www.poynter.org/ifcn/anti-misinformation-actions/>.
- Furnémont, Jean-François, and Deirdre, Kevin. *Regulation of Political Advertising: a comparative study with reflections on the situation in South-East Europe*. Council of Europe and the European Union, September 2020. <https://rm.coe.int/study-on-political-advertising-eng-final/1680a0c6e0>
- Gadde, Vijaya and Beykpour, Kayvon. "Additional steps we're taking ahead of the 2020 US Election." *Twitter*, October 9, 2020. [https://blog.twitter.com/en\\_us/topics/company/2020/2020-election-changes.html](https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html).

- Gebhart, Gennie. "Who Has Your Back? Censorship Edition 2019." *Electronic Frontier Foundation (EFF)*, June 12, 2019. <https://www.eff.org/wp/who-has-your-back-2019>.
- Gingras, Richard. "Labeling fact-check articles in Google News." *Google News Initiative*, October 13, 2016. <https://blog.google/outreach-initiatives/google-news-initiative/labeling-fact-check-articles-google-news/>.
- Glaser, April. "Google is rolling out a fact-check feature in its search and news results." *Recode*, April 8, 2017. <https://www.vox.com/2017/4/8/15229878/google-fact-check-fake-news-search-news-results>.
- Gleicher, Nathaniel. "Labeling State-Controlled Media On Facebook." *Facebook*, June 4, 2020. <https://about.fb.com/news/2020/06/labeling-state-controlled-media/>.
- Global Internet Forum to Counter Terrorism. "Governance." Accessed on May 11, 2021. <https://gifct.org/governance/>.
- Global Internet Forum to Counter Terrorism. "Joint Tech Innovation." Accessed on May 11, 2021. <https://gifct.org/joint-tech-innovation/>.
- Global Internet Forum to Counter Terrorism. "Membership." Accessed on May 11, 2021. <https://gifct.org/membership/>.
- Global Internet Forum to Counter Terrorism. "Transparency." Accessed on May 11, 2021. <https://gifct.org/transparency/>.
- Global Internet Forum to Counter Terrorism. "What is the hash sharing consortium and how does it work?" Accessed May 11, 2021. <https://gifct.org/?faqs=what-is-the-hash-sharing-consortium-and-how-does-it-work>.
- Gomes, Ben. "Our latest quality improvements for Search." *Google*, April 25, 2017. <https://www.blog.google/products/search/our-latest-quality-improvements-search/>.
- Global Network on Extremism and Technology. Accessed on May 11, 2021. <https://gnet-research.org/>.
- Google. "About knowledge panels." Accessed May 12, 2021. <https://support.google.com/knowledgepanel/answer/9163198?hl=en>.
- Google. "COVID-19." Accessed on May 12, 2021. <https://www.google.com/search?q=covid+19&oq=covid+19&aqs=edge.O0j69i60j69i61j69i60j0l3.1874j0j1&sourceid=chrome&ie=UTF-8>.
- Google. *EC EU Code of Practice on Disinformation: Google Annual Report*. Google: 2019.
- Google. *EU & COVID-19 Disinformation Google Report, January 2021*. Google: 2021.
- Google. *EU Code of Practice on Disinformation: Google Report*. 2019. [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-5/google\\_-\\_ec\\_action\\_plan\\_reporting\\_CF162236-E8FB-725E-C0A3D2D6CCFE678A\\_56994.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-5/google_-_ec_action_plan_reporting_CF162236-E8FB-725E-C0A3D2D6CCFE678A_56994.pdf).
- Google. "Fact Check." Last updated: March 18, 2021. <https://developers.google.com/search/docs/data-types/factcheck>.
- Google. "Fact Check Explorer." Accessed May 12, 2021. <https://toolbox.google.com/factcheck/explorer>.
- Google. "Featured Policies." *Google Transparency Report*, Accessed on May 11, 2021. <https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism>.
- Google. "Google Help Communities Content Policy." Accessed May 12, 2021. <https://support.google.com/communities/answer/7425194?hl=en>.
- Google. "Google key milestones for the Implementation of the Code of Practice on Disinformation."
- Google. "Help improve Google's products." Accessed on May 11, 2021. <https://www.google.com/tools/feedback/intl/en/>.
- Google. *How Google Fights Disinformation*. Google: 2019. [https://www.blog.google/documents/37/How\\_Google\\_Fights\\_Disinformation.pdf](https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf).
- Google. *How Google Fights Piracy*. Google: 2018. [https://www.blog.google/documents/25/GO806\\_Google\\_FightsPiracy\\_eReader\\_final.pdf](https://www.blog.google/documents/25/GO806_Google_FightsPiracy_eReader_final.pdf).

- Google. "How Search algorithms work." Accessed May 12, 2021. <https://www.google.com/search/howsearchworks/algorithms/>.
- Google. "Privacy and Terms." March 31, 2020. <https://policies.google.com/terms?hl=en>.
- Google. "Threat Analysis Group." Accessed May 12, 2021. <https://blog.google/threat-analysis-group>.
- Gorwa, Robert; Binns, Reuben; and Katzenbach, Christian. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." *Big Data and Society* (January-June 2020): 1-15. <https://doi.org/10.1177%2F2053951719897945>.
- Gorwa, Robert. "The platform governance triangle: conceptualising the informal regulation of online content." *Internet Policy Review* 8, no. 2 (June 2019). DOI: 10.14763/2019.2.1407.
- Gorwa, Robert. "Regulating them softly." *Centre for International Governance Innovation*, October 28, 2019. <https://www.cigionline.org/articles/regulating-them-softly>.
- Government of Canada. *Charlevoix commitment on defending democracy from foreign threats*. Government of Canada, 2019. [https://www.international.gc.ca/world-monde/international\\_relations-relations\\_internationales/g7/documents/2018-06-09-defending\\_democracy-defense\\_democratie.aspx?lang=eng](https://www.international.gc.ca/world-monde/international_relations-relations_internationales/g7/documents/2018-06-09-defending_democracy-defense_democratie.aspx?lang=eng)
- Grassegger, Hannes and Krogerus, Mikael. "Fake news and botnets: how Russia weaponised the web." *The Guardian*, December 2, 2017. <https://www.theguardian.com/technology/2017/dec/02/fake-news-botnets-how-russia-weaponised-the-web-cyber-attack-estonia>.
- Greenberg, Julia. "Why Facebook and Twitter Can't Just Wipe Out ISIS Online." *Wired*, November 21, 2015. <https://www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/>.
- Grimme, Christian; Assenmacher, Dennis; and Adam, Lena. "Changing Perspectives: Is it Sufficient to Detect Social Bots?" *Social Computing and Social Media. User Experience and Behavior* (2018): 445-461. [https://link.springer.com/chapter/10.1007/978-3-319-91521-0\\_32](https://link.springer.com/chapter/10.1007/978-3-319-91521-0_32).
- Gross, Grant. "After WCIT, US lawmakers look for ways to advance Internet freedom." *Computerworld*, February 5, 2013. <https://www.computerworld.com/article/2494615/after-wcit--us-lawmakers-look-for-ways-to-advance-internet-freedom.html>.
- Gross, Grant. "World telecom conference ends with uneven support." *PCWorld*, December 15, 2012. <https://www.pcworld.com/article/2020583/world-telecom-conference-ends-with-uneven-support.html/>
- Haan, Sarah C. "Bad Actors: Authenticity, Inauthenticity, Speech, and Capitalism." *Journal of Constitutional Law* 22, no. 3 (May 2020): 619-686. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3458795](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458795).
- Hadavas, Chloe. "The Future of Free Speech Online May Depend on This Database." *Slate*, August 13, 2020. <https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html>.
- Hakeem, Rawan. "Highlighting the diversity of content in Google News." *Google*, September 17, 2009. <https://news.googleblog.com/2009/09/highlighting-diversity-of-content-in.html>.
- Hao, Karen. "How Facebook got addicted to spreading misinformation." *MIT Technology Review*, March 11, 2021. <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>.
- Harvard Kennedy School. "Tackling misinformation: What researchers could do with social media data." *HKS Misinformation Review*, December 9, 2020. <https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/>.
- Hatmaker, Taylor. "Facebook Oversight Board says other social networks 'welcome to join' if project succeeds." *Techcrunch*, February 11, 2021. <https://techcrunch.com/2021/02/11/facebook-oversight-board-other-social-networks-beyond-facebook/>.
- Hauser, Christine. "GoDaddy Severs Ties With Daily Stormer After Charlottesville Article." *The New York Times*, August 14, 2017. <https://www.nytimes.com/2017/08/14/us/godaddy-daily-stormer-white-supremacists.html>.



- Heath, Alex. "Facebook is going to use Snopes and other fact-checkers to combat and bury 'fake news'." *Business Insider*, December 15, 2016. <https://www.businessinsider.nl/facebook-will-fact-check-label-fake-news-in-news-feed-2016-12?international=true&r=US>.
- Hernández, Gina. "TikTok añade nuevas indicaciones que ayudan a reconsiderar antes de compartir." *TikTok*, accessed May 11, 2021. <https://tiktok.prezly.com/tiktok-anade-nuevas-indicaciones-que-ayudan-a-reconsiderar-antes-de-compartir>.
- Hirsch, Dennis D. "The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-Regulation?" *Seattle University Law Review* 34, no. 2 (2011): 439-480. <https://ssrn.com/abstract=1758078>.
- Hollis, Duncan B. "The Influence of War; The War for Influence." *Temple International & Comparative Law Journal* 32, no. 1 (2018). [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3155273](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3155273).
- Home Office (UK); Department for Education (UK). *How social media is used to encourage travel to Syria and Iraq: Briefing note for schools*. UK: 2015. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/440450/How\\_social\\_media\\_is\\_used\\_to\\_encourage\\_travel\\_to\\_Syria\\_and\\_Iraq.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/440450/How_social_media_is_used_to_encourage_travel_to_Syria_and_Iraq.pdf).
- Howell, Jen Patja. "The Lawfare Podcast: Alex Stamos on Fighting Election Disinformation in Real Time." *Lawfare*, August 20, 2020. <https://www.lawfareblog.com/lawfare-podcast-alex-stamos-fighting-election-disinformation-real-time>.
- Howell, Jen Patja. "The Lawfare Podcast: Collaborating to Counter Violent Extremism Online." *Lawfare*, November 25, 2020. <https://www.lawfareblog.com/lawfare-podcast-collaborating-counter-violent-extremism-online>.
- Huseinzade, Nazrin. "Algorithm Transparency: How to Eat the Cake and Have It Too." *European Law Blog*, January 27, 2021. <https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/>.
- International Court of Justice. "Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)." 1986. <https://www.icj-cij.org/public/files/case-related/70/070-19860627-JUD-01-00-EN.pdf>.
- International Fact-Checking Network. "IFCN releases a statement about accusations against one of its verified signatories." *Poynter*, September 11, 2019. <https://www.poynter.org/fact-checking/2019/ifcn-releases-a-statement-about-accusations-against-one-of-its-verified-signatories/>.
- Internet Advisory Board. "Standards, Guidelines & Best Practices." Accessed on May 11, 2021. [https://www.iab.com/guidelines/?post\\_type=iab\\_guideline](https://www.iab.com/guidelines/?post_type=iab_guideline).
- Isaac, Mike. "Facebook Ends Ban on Political Advertising." *New York Times*, March 3, 2021. <https://www.nytimes.com/2021/03/03/technology/facebook-ends-ban-on-political-advertising.html>.
- Jardine, Eric. "Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet." *First Monday* 24, no. 12 (December 2019). DOI: <http://dx.doi.org/10.5210/fm.v24i12.10266>.
- No Author. "Justice Collaboratory to Lead Facebook Data Transparency Advisory Group." *Yale Law School*, October 2, 2018. <https://law.yale.edu/yls-today/news/justice-collaboratory-lead-facebook-data-transparency-advisory-group>.
- Kampanakis, Panos. "Security Automation and Threat Information-Sharing Options." *IEEE Security & Privacy* 12, no. 5 (2014): 42-51. DOI: 10.1109/MSP.2014.99.
- Katzenstein, Peter J. *The Culture of National Security: Norms and Identity in World Politics*. New York: Columbia University Press, 1996.
- Kayali, Laura. "TikTok launches 'Safety Advisory Council' in Europe." *Politico*, March 2, 2021. <https://www.politico.eu/article/tiktok-launches-safety-advisory-council-in-europe/>.
- Kaye, David. *Speech Police: The Struggle to Govern the Internet*. New York: Columbia Global Reports, 2019.
- Keohane, Robert. "Social Norms and Agency in World Politics." *NYU School of Law* 14 (2010). <http://www.law.nyu.edu/sites/default/files/siwp/Keohane.pdf>.



- Khan, Irene. "Disinformation and freedom of opinion. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Kahn". *United Nations*: April 13, 2021. <https://undocs.org/A/HRC/47/25>
- Kimball, Whitney. "TikTok Is Adding a Potential Misinformation Warning Label to Save Us From Ourselves." *Gizmodo*, February 3, 2021. <https://gizmodo.com/tiktok-is-adding-a-potential-misinformation-warning-lab-1846189941>.
- Klimburg, Alexander and Almeida, Virgilio. "Cyber Peace and Cyber Stability: Taking the Norm Road to Stability." *IEEE Internet Computing* 23, no. 4 (July-Aug. 2019): 61-66. <https://ieeexplore.ieee.org/document/8874985>.
- Klimburg, Alexander and Faesen, Louk. "A Balance of Power in Cyberspace." In *Governing Cyberspace: Behavior, Power, and Diplomacy*, edited by Dennis Broeders and Bibi van den Berg. London: Rowman & Littlefield, 2020.
- Klimburg, Alexander. *The Darkening Web: The War for Cyberspace*. New York: Penguin Books, 2017.
- Klimburg, Alexander. "The Internet Yalta." *Center for a New American Security*. February 5, 2013. <https://www.jstor.org/stable/resrep06186>.
- Klimburg, Alexander (Ed.). *National Cyber Security Framework Manual*. Tallinn: NATO CCD COE Publications, 2012. [https://www.ccdcoe.org/uploads/2018/10/NCSFM\\_0.pdf](https://www.ccdcoe.org/uploads/2018/10/NCSFM_0.pdf).
- Knake, Robert K. "At Facebook, One Million Takedowns Per Day Is Evidence of Failure, Not Success." *Council on Foreign Relations*, February 20, 2020. <https://www.cfr.org/blog/facebook-one-million-takedowns-day-evidence-failure-not-success>.
- Knake, Robert K. "Banning Covert Foreign Election Interference." *Council on Foreign Relations*, May 29, 2020. [https://www.cfr.org/report/banning-covert-foreign-election-interference?utm\\_medium=social\\_share&utm\\_source=tw](https://www.cfr.org/report/banning-covert-foreign-election-interference?utm_medium=social_share&utm_source=tw).
- Knuutila, Aleks; Herasimenka, Aliaksandr; Au, Hubert; Bright, Jonathan; Nielsen, Rasmus; and Howard, Philip N. "COVID-related Misinformation on YouTube: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies." *Oxford Internet Institute* (2020). <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/09/YouTube-misinfo-memo.pdf>.
- Kosslyn, Justin and Yu, Cong. "Fact Check now available in Google Search and News around the world." *Facebook*, April 7, 2017. <https://blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/>.
- Krasner, Stephen D. "Structural Causes and Regime Consequences: Regimes as Intervening Variables." *International Organization* 36, no. 2 (Spring 1982): 185-205. <https://www.jstor.org/stable/2706520>.
- Lada, Akos; Wang, Meihong; and Yan, Tak. "How does News Feed predict what you want to see?" *Facebook*, January 26, 2021. <https://tech.fb.com/news-feed-ranking/>.
- Landi, Martyn. "TikTok adds new vaccine misinformation labels and strengthens community rules." *Breakingnews.ie*, December 15, 2020. <https://www.breakingnews.ie/business/tiktok-adds-new-vaccine-misinformation-labels-and-strengthens-community-rules-1051447.html>.
- Lau, Tim. "The Honest Ads Act Explained." *The Brennan Center for Justice*, January 17, 2020. <https://www.brennancenter.org/our-work/research-reports/honest-ads-act-explained>.
- Law Library of Congress. *Initiatives to Counter Fake News in Selected Countries*. The Law Library of Congress: April 2019. <https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf>.
- Lee, Dave. "Key fact-checkers stop working with Facebook." *BBC*, February 2, 2019. <https://www.bbc.com/news/technology-47098021>.
- Lerman, Rachel. "Parler is back online, more than a month after tangle with Amazon knocked it offline." *The Washington Post*, February 15, 2021. <https://www.washingtonpost.com/technology/2021/02/15/parler-returns-online/>.
- Lerman, Rachel. "Seeing isn't always believing: Google starts fact-checking images." *The Washington Post*, June 23, 2020. <https://www.washingtonpost.com/technology/2020/06/22/google-fact-check-images/>.

- Lever, Rob. "Fake Facebook accounts: Never-ending battle against bots." *The Jakarta Post*, May 25, 2019. <https://www.thejakartapost.com/life/2019/05/25/fake-facebook-accounts-never-ending-battle-against-bots.html>.
- Levin, Sam. "'They don't care': Facebook factchecking in disarray as journalists push to cut ties." *The Guardian*, December 13, 2018. <https://www.theguardian.com/technology/2018/dec/13/they-dont-care-facebook-fact-checking-in-disarray-as-journalists-push-to-cut-ties>.
- Lewis, James A. "Liberty, Equality, Connectivity: Transatlantic Cybersecurity Norms." *Center for Strategic & International Studies*, February 2014. [https://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy\\_files/files/publication/140225\\_Lewis\\_TransatlanticCybersecurityNorms.pdf](https://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/140225_Lewis_TransatlanticCybersecurityNorms.pdf).
- LinkedIn Help. "Recognize and Report Spam, Inappropriate, and Abusive Content." *LinkedIn*, last updated April 2021. <https://www.linkedin.com/help/linkedin/answer/37822>.
- Llansó, Emma. "Platforms Want Centralized Censorship. That Should Scare You." *Wired*, April 18, 2019. <https://www.wired.com/story/platforms-centralized-censorship/>.
- Lyons, Tessa. "Hard Questions: What's Facebook's Strategy for Stopping False News?" *Facebook*, May 23, 2018. <https://about.fb.com/news/2018/05/hard-questions-false-news/>.
- Mair, David. "#Westgate: A Case Study: How al-Shabaab used Twitter during an Ongoing Attack." *Studies in Conflict and Terrorism* 40, no. 1 (2017): 24–43. <https://doi.org/10.1080/1057610X.2016.1157404>.
- Marr, Bernard. "Coronavirus Fake News: How Facebook, Twitter, And Instagram Are Tackling The Problem." *Forbes*, March 27, 2020. <https://www.forbes.com/sites/bernardmarr/2020/03/27/finding-the-truth-about-covid-19-how-facebook-twitter-and-instagram-are-tackling-fake-news/?sh=430a82919771>.
- Marsden, Christopher. *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*. Cambridge: Cambridge University Press, 2011.
- Marsden, Chris; Meyer, Trisha; and Brown, Ian. "Platform values and democratic elections: How can the law regulate digital disinformation?" *Computer Law & Security Review* 36 (2020). <https://doi.org/10.1016/j.clsr.2019.105373>.
- Masnick, Mike. "Who Signed The ITU WCIT Treaty... And Who Didn't." *Techdirt*, December 14, 2012. <https://www.techdirt.com/articles/20121214/14133321389/who-signed-itu-wcit-treaty-who-didnt.shtml>.
- McFaul, Michael. *Securing American Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond*. Stanford: Stanford Cyber Policy Center, June 2019. <https://fsi.stanford.edu/publication/securing-american-elections-prescriptions-enhancing-integrity-and-independence-2020-us>.
- McSherry, Corynne; McKinney, India; York, Jillian C. "Content Moderation Is A Losing Battle. Infrastructure Companies Should Refuse to Join the Fight." *Electronic Frontier Foundation*, April 1, 2021. <https://www.eff.org/deeplinks/2021/04/content-moderation-losing-battle-infrastructure-companies-should-refuse-join-fight>.
- Meserole, Chris and Byman, Daniel. *Terrorist Definitions and Designations Lists: What Technology Companies Need to Know*. London: RUSI, 2019. <https://www.brookings.edu/wp-content/uploads/2019/07/GRNTT-Paper-No.-7.pdf>.
- Microsoft. "EU Code of Practice on Disinformation." 2019. [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-24/microsoft\\_eu\\_cop\\_on\\_disinformation\\_signatory\\_document\\_may\\_2019\\_asd\\_final\\_53FCB030-02E6-C461-9ACDA79C0BCCEAAB\\_60091.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-24/microsoft_eu_cop_on_disinformation_signatory_document_may_2019_asd_final_53FCB030-02E6-C461-9ACDA79C0BCCEAAB_60091.pdf).
- Microsoft. *January Update on Microsoft Corporation's Efforts to Tackle COVID-19 Disinformation*. Microsoft: 2021.
- Microsoft. "Microsoft Community Code of Conduct." Accessed May 12, 2021. <https://answers.microsoft.com/en-us/page/codeofconduct>.
- Microsoft. *Microsoft Self-Assessment and Report on Compliance with the EU Code of Practice on Disinformation*. Microsoft: 2019.
- Microsoft. "Microsoft Services Agreement." August 1, 2020. <https://www.microsoft.com/en/servicesagreement/>.

- Microsoft Bing Blogs. "Bing adds Fact Check label in SERP to support the ClaimReview markup." *Microsoft*, September 14, 2017. <https://blogs.bing.com/Webmaster-Blog/September-2017/Bing-adds-Fact-Check-label-in-SERP-to-support-the-ClaimReview-markup#content>.
- Microsoft Corporate Blogs. "Global Internet Forum to Counter Terrorism has first meeting Aug. 1." *Microsoft*, July 31, 2017. <https://blogs.microsoft.com/on-the-issues/2017/07/31/global-internet-forum-counter-terrorism-first-meeting-aug-1/>.
- Miller, Maggie. "Social media bots pose threat ahead of 2020." *The Hill*, August 6, 2019. <https://thehill.com/policy/cybersecurity/456282-social-media-bots-pose-threat-ahead-of-2020>.
- Ministry of Defense France. "International Law Applied to Operations in Cyberspace." Ministry of Defense: 2019. <https://www.defense.gouv.fr/content/download/567648/9770527/file/international+law+applied+to+operations+in+cyberspace.pdf#page=6>.
- The Ministry of Foreign Affairs of the Russian Federation. "Convention On International Information Security." September 22, 2011. [https://www.mid.ru/en/foreign\\_policy/official\\_documents/-/asset\\_publisher/CptlCk6B6BZ29/content/id/191666](https://www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptlCk6B6BZ29/content/id/191666).
- Ministry of Foreign Affairs of the Russian Federation. *Doctrine of Information Security of the Russian Federation*. No. 646. Moscow: Ministry of Foreign Affairs of Russia, 2016. [https://www.mid.ru/en/foreign\\_policy/official\\_documents/-/asset\\_publisher/CptlCk6B6BZ29/content/id/2563163](https://www.mid.ru/en/foreign_policy/official_documents/-/asset_publisher/CptlCk6B6BZ29/content/id/2563163)
- Monster, Rob. "Why Epik welcomed Gab.com." *Epik*, November 3, 2018. <https://www.epik.com/blog/why-epik-welcomed-gab-com.html>.
- Morgan, Kevin. "Taking action against COVID-19 vaccine misinformation." *Twitter*, December 15, 2020. <https://newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation>.
- Morgus, Robert. "Russia Gains an Upper Hand in the Cyber Norms Debate". *Council on Foreign Relations*, December 5, 2016. <https://www.cfr.org/blog/russia-gains-upper-hand-cyber-norms-debate>
- Morris, Lyle J.; Mazarr, Michael J.; Hornung, Jeffery W.; Pezard, Stephanie; Binnendijk, Anika; Keep, Marta. *Gaining Competitive Advantage in the Gray Zone: Response Options for Coercive Aggression Below the Threshold of Major War*. Santa Monica: RAND Corporation, 2019. [https://www.rand.org/pubs/research\\_reports/RR2942.html](https://www.rand.org/pubs/research_reports/RR2942.html).
- Morse, Jack. "Why social media companies won't kill off bots." *Mashable*, February 7, 2018. <https://mashable.com/2018/02/06/facebook-instagram-twitter-bots/?europa=true#text=Facebook%20isn%27t%20much%20different.&text=The%20more%20they%20engage%20with,contrary%2C%20it%20thrives%20on%20them>.
- Mosseri, Adam. "Addressing Hoaxes and Fake News." *Facebook*, December 15, 2016. <https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/>.
- Moynihan, Harriet. "Deplatforming Trump puts big tech under fresh scrutiny." *Chatham House*, January 22, 2021. <https://www.chathamhouse.org/2021/01/deplatforming-trump-puts-big-tech-under-fresh-scrutiny>.
- Mozilla. *Update on Milestones for the Implementation of the Code of Practice on Disinformation*. Belgium: Mozilla, 2019. [https://ec.europa.eu/information\\_society/newsroom/image/document/2019-5/mozilla\\_cop\\_report\\_-\\_18\\_01\\_19\\_CF162508-CF98-8ACD-89BCCC1BA4230DD9\\_56995.pdf](https://ec.europa.eu/information_society/newsroom/image/document/2019-5/mozilla_cop_report_-_18_01_19_CF162508-CF98-8ACD-89BCCC1BA4230DD9_56995.pdf).
- Murray, Alex. "How to report fake news to social media." *BBC*, November 22, 2016. <https://www.bbc.com/news/38053324>.
- Nassetta, Jack and Gross, Kimberly. "State media warning labels can counteract the effects of foreign misinformation." *Harvard Kennedy School (HKS) Misinformation Review* 1, no. 7 (October 2020). DOI: <https://doi.org/10.37016/mr-2020-45>.
- The Netherlands. "The Kingdom of the Netherlands' response to the pre-draft report of the OEWG." 2020. <https://front.un-arm.org/wp-content/uploads/2020/04/kingdom-of-the-netherlands-response-pre-draft-oweg.pdf>.
- <https://www.mfat.govt.nz/en/peace-rights-and-security/international-security/christchurch-call/>.
- NewsGuard. "Microsoft Expands NewsGuard Adoption." May 14, 2020. <https://www.newsguardtech.com/press/microsoft-expands-newsguard-adoption/>.

- No Author. "Regulate online political ads for greater political integrity." Transparency International. March 10, 2021. <https://www.transparency.org/en/news/regulate-online-political-ads-for-greater-political-integrity>
- Nyhan, Brendan and Reifler, Jason. "When Corrections Fail: The persistence of political misperceptions." *Political Behavior* 32 (March 2010): 303-330. <https://link.springer.com/article/10.1007/s11109-010-9112-2>.
- Ohlin, Jens David. "Election Interference: The Real Harm and The Only Solution". *Cornell Legal Studies Research Paper*, No. 18-50 (2018). <https://ssrn.com/abstract=3276940>.
- Ohlin, Jens David. *Election Interference: International Law and the Future of Democracy*. Cambridge: Cambridge University Press, 2020. doi:10.1017/9781108859561
- Ohlin, Jens David. "Did Russian Cyber Interference in the 2016 Election Violate International Law?" *Texas Law Review* 95 (2017): 1579-1598. <https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2632&context=facpub>.
- Ong, Thuy. "YouTube will start labeling videos from state-funded broadcasters." *The Verge*, February 2, 2018. <https://www.theverge.com/2018/2/2/16964190/youtube-state-funded-broadcasters>.
- Ördén, Hedvig, and James Pamment. *What Is So Foreign About Foreign Influence Operations?*. Washington DC: Carnegie Endowment for International Peace, 2021. [https://carnegieendowment.org/files/Orden\\_Pamment\\_ForeignInfluenceOps2.pdf](https://carnegieendowment.org/files/Orden_Pamment_ForeignInfluenceOps2.pdf)
- Oversight Board. Accessed May 12, 2021. <https://oversightboard.com/>.
- Oversight Board. "Case Decision 2020-006-FB-FBR." Last Modified: January 28, 2021. <https://www.oversightboard.com/decision/FB-XWJQBU9A/>.
- Oversight Board. *Oversight Board Bylaws*. Oversight Board: 2021. <https://www.oversightboard.com/sr/governance/bylaws>.
- Oversight Board. "Oversight Board upholds former President Trump's suspension, finds Facebook failed to impose proper penalty." May 2021. <https://oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty/>.
- Oversight Board. *Rulebook for Case Review and Policy Guidance*. Oversight Board: 2020. <https://oversightboard.com/sr/rulebook-for-case-review-and-policy-guidance>.
- Pallero, Javier. *Protecting Free Expression in the Era of Online Content Moderation: Access Now's preliminary recommendations on content moderation and Facebook's planned oversight board*. AccessNow: 2019. <https://www.accessnow.org/cms/assets/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf>.
- Pamment, James. *EU Code of Practice on Disinformation: Briefing Note for the New European Commission*. Washington, DC: the Carnegie Endowment for International Peace, 2020. [https://carnegieendowment.org/files/Pamment\\_-\\_EU\\_Code\\_of\\_Practice.pdf](https://carnegieendowment.org/files/Pamment_-_EU_Code_of_Practice.pdf).
- Pamment, James. *The EU's Role in Fighting Disinformation: Taking Back the Initiative*. Washington DC: the Carnegie Endowment for International Peace, 2020. [https://carnegieendowment.org/files/Pamment\\_-\\_Future\\_Threats.pdf](https://carnegieendowment.org/files/Pamment_-_Future_Threats.pdf).
- Pappas, Vanessa. "Combating misinformation and election interference on TikTok." *TikTok*, August 5, 2020. <https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok>.
- Pasternack, Alex. "Facebook quietly pressured its fact-checkers over climate and abortion posts." *Fast Company*, August 20, 2020. <https://www.fastcompany.com/90538655/facebook-is-quietly-pressuring-its-independent-fact-checkers-to-change-their-rulings>.
- Paul, Kari; Harding, Luke; and Carrell, Severin. "Far-right website 8kun again loses internet service protection following Capitol attack." *The Guardian*, January 15, 2021. <https://www.theguardian.com/technology/2021/jan/15/8kun-8chan-capitol-breach-violence-isp>.
- Pennycook, Gordon; Bear, Adam; Collins, Evan T.; Rand, David G. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without

- Warnings." *Management Science* 66, no. 11 (November 2020): 4944–4957. <https://doi.org/10.1287/mnsc.2019.3478>.
- Plasilova, Iva; Hill, Jordan; Carlberg, Malin; Goubet, Marion; and Procee, Richard. *Study for the "Assessment of the implementation of the Code of Practice on Disinformation"*. Luxembourg: Publications Office of the European Union, 2020. <https://digital-strategy.ec.europa.eu/en/library/study-assessment-implementation-code-practice-disinformation>.
- Plumb, Radha Iyengar. "An Independent Report on How We Measure Content Moderation." *Facebook*, May 23, 2019. <https://about.fb.com/news/2019/05/dtag-report/>.
- Politi, Daniel. "Facebook Pushes Back After Biden Accused It of "Killing People" With Disinformation." *Slate*. July 18, 2021. <https://slate.com/news-and-politics/2021/07/facebook-pushes-back-biden-killing-people-misinformation.html>
- Poynter. "The commitments of the code of principles." *ICFN Code of Principles*, accessed May 6, 2021. <https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles>.
- Prince, Matthew. "Terminating Service for 8Chan." *Cloudflare*, August 5, 2019. <https://blog.cloudflare.com/terminating-service-for-8chan/>.
- Prince, Matthew. "Why We Terminated Daily Stormer." *Cloudflare*, August 17, 2017. <https://blog.cloudflare.com/why-we-terminated-daily-stormer/>.
- Rainie, Lee; Anderson, Jenna; and Albright, Jonathan. "The Future of Free Speech, Trolls, Anonymity and Fake News Online." *Pew Research Center: Internet and Technology*, March 29, 2017. <https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>.
- RAND. "Schema.org Claim Review." Accessed May 11, 2021. <https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search/items/schemaorg-claim-review.html>.
- Reddit. "Mod Support." Accessed May 11, 2021. <https://www.reddit.com/r/ModSupport/>.
- Robertson, Adi. "Facebook Oversight Board overturns hate speech and pandemic misinformation takedowns." *The Verge*, January 28, 2021. <https://www.theverge.com/2021/1/28/22254155/facebook-oversight-board-first-rulings-coronavirus-misinformation-hate-speech>.
- Robertson, Adi. "Facebook starts labeling 'state-controlled media' pages." *The Verge*, June 4, 2020. <https://www.theverge.com/2020/6/4/21280542/facebook-state-controlled-media-account-post-label-election-interference-ads-rt>.
- Roguski, Przemyslaw. "The Importance of New Statements on Sovereignty in Cyberspace by Austria, the Czech Republic and United States." *Just Security*, May 11, 2020. <https://www.justsecurity.org/70108/the-importance-of-new-statements-on-sovereignty-in-cyberspace-by-austria-the-czech-republic-and-united-states/>.
- Rõigas, Henry. "An Updated Draft of the Code of Conduct Distributed in the United Nations – What's New?" *CCDCOE*, 2015. <https://ccdcoe.org/incyde-articles/an-updated-draft-of-the-code-of-conduct-distributed-in-the-united-nations-whats-new/>.
- Rõigas, Henry and Minárik, Tomáš. "2015 UN GGE Report: Major Players Recommending Norms of Behaviour, Highlighting Aspects of International Law." *CCDOE*, 2015. <https://ccdcoe.org/incyde-articles/2015-un-gge-report-major-players-recommending-norms-of-behaviour-highlighting-aspects-of-international-law/>
- Rosen, Guy; Harbath, Katie; Gleicher, Nathaniel; and Leathern, Rob. "Helping to Protect the 2020 US Elections." *Facebook*, October 21, 2019. <https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/>.
- Rosen, Guy. "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19." *Facebook*, April 16, 2020. <https://about.fb.com/news/2020/04/covid-19-misinfo-update/>.
- Roth, Yoel and Achuthan, Ashita. "Building rules in public: Our approach to synthetic & manipulated media." *Twitter*, February 4, 2020. [https://blog.twitter.com/en\\_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html](https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html).
- Roth, Yoel and Pickles, Nick. "Updating our approach to misleading information." *Twitter*, May 11, 2020. [https://blog.twitter.com/en\\_us/topics/product/2020/updating-our-approach-to-misleading-information.html](https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html).



- Russia. "Statement by President of Russia Vladimir Putin on a comprehensive program of measures for restoring the Russia – US cooperation in the filed [sic] of international information security." September 25, 2020. <http://en.kremlin.ru/events/president/news/64086>.
- Saltz, Emily; Leibowicz, Claire; and Wardle, Claire. "Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions." (December 2020). <https://arxiv.org/pdf/2011.12758.pdf>.
- Saltz, Emily; Shane, Tommy; Kwan, Victoria; Leibowicz, Claire; and Wardle, Claire. "It matters how platforms label manipulated media. Here are 12 principles designers should follow." *Partnership on AI*, June 9, 2020. <https://www.partnershiponai.org/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/>.
- Samek, Geoff. "Greater transparency for users around news broadcasters." *YouTube*, February 2, 2018. <https://blog.youtube/news-and-events/greater-transparency-for-users-around>.
- Schema.org. "ClaimReview." Last updated March 8, 2021. <https://schema.org/ClaimReview>.
- Schmitt, Michael. "The Defense Department's Measured Take on International Law in Cyberspace." *Just Security*, March 11, 2020. <https://www.justsecurity.org/69119/the-defense-departments-measured-take-on-international-law-in-cyberspace/>.
- Schmitt, Michael. "'Virtual' Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law." *Chicago Journal of International Law* 19, no. 1 (2018): 30–67. <https://chicagounbound.uchicago.edu/cjil/vol19/iss1/2>
- Scott, Mark. "Twitter Fails E.U. Standard on Removing Hate Speech Online." *New York Times*, May 31, 2017. <https://www.nytimes.com/2017/05/31/technology/twitter-facebook-google-europe-hate-speech.html>.
- Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department. *Online Harms White Paper*. UK: HM Government, 2019. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/973939/Online\\_Harms\\_White\\_Paper\\_V2.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf).
- Seetharaman, Deepa and Horwitz, Jeff. "Facebook Creates Teams to Study Racial Bias, After Previously Limiting Such Efforts." *The Wall Street Journal*, July 21, 2020. [https://www.wsj.com/articles/facebook-creates-teams-to-study-racial-bias-on-its-platforms-11595362939?mod=hp\\_lista\\_pos1](https://www.wsj.com/articles/facebook-creates-teams-to-study-racial-bias-on-its-platforms-11595362939?mod=hp_lista_pos1).
- Sehl, Katie. "How the Twitter Algorithm Works in 2020 and How to Make it Work for You." *Hootsuite*, May 20, 2020. <https://blog.hootsuite.com/twitter-algorithm/#:~:text=Twitter%27s%20algorithm%2C%20like%20most%20social,algorithms%2C%20is%20all%20about%20personalization.&text=All%20social%20algorithms%20use%20machine,rich%20media%2C%20and%20other%20factors>.
- Shah, Syed Ali Raza Shah and Issac, Biju. "Performance Comparison of Intrusion Detection Systems and Application of Machine Learning to Snort System." *Future Generation Computer Systems* 80 (March 2018): 157–70. <https://doi.org/10.1016/j.future.2017.10.016>.
- Shane, Scott. "Russia Isn't the Only One Meddling in Elections. We Do It, Too." *The New York Times*, February 17, 2018. <https://www.nytimes.com/2018/02/17/sunday-review/russia-isnt-the-only-one-meddling-in-elections-we-do-it-too.html>.
- Shao, Chengcheng; Ciampaglia, Giovanni Luca; Varol, Onur; Yang, Kai-Cheng; Flammini, Alessandro; and Menczer, Filippo. "The spread of low-credibility content by social bots." *Nature Communications* 9, no. 4787 (2018). DOI: <https://dx.doi.org/10.1038%2Fs41467-018-06930-7>.
- Shields, Mike. "Snap suddenly has a leg up on Facebook and Google — but it still needs to do 2 things to steal their advertisers." *Business Insider*, October 7, 2017. <https://www.businessinsider.com/snapchats-closed-doors-keep-fake-news-out-2017-10?international=true&r=US&IR=T>.
- Shimer, David. *Rigged: America, Russia and 100 Years of Covert Electoral Interference*. London: William Collins, 2020.
- Shu, Catherine. "Facebook will ditch Disputed Flags on fake news and display links to trustworthy articles instead." *Techcrunch*, December 21, 2017. <https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/>.



- Silverman, Craig. "Facebook Removed Over 2 Billion Fake Accounts, But The Problem Is Getting Worse." *Buzzfeed*, May 24, 2019. <https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-accounts-afd>.
- Singh, Spandana and Bagchi, Koustubh. "How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19: Reddit." *New America*, June 1, 2020. <https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/>.
- Smith, Brad; Donfried, Karen; and LeBlanc, Dominic. *Multi-Stakeholder Insights: A Compendium on Countering Election Interference*. Canada: 2021. <https://www.canada.ca/content/dam/di-id/documents/compendium-eng.pdf>.
- Social Media Today. "Keeping up with the Algorithms." Accessed May 11, 2021. <https://www.socialmedia-today.com/topic/algorithm-updates/>.
- Softness, Nicole. "Terrorist Communications: Are Facebook, Twitter, and Google Responsible for the Islamic State's Actions?" *Journal of International Affairs* 70, no. 1 (Winter 2016): 201-215. <https://www.jstor.org/stable/90012606>.
- The Sounding Board. "The Sounding Board's unanimous final opinion on the so-called code of practice." EBU, September 28, 2018. <https://www.ebu.ch/news/2018/09/sounding-board-of-forum-on-disinformation-online-issues-unanimous-opinion-on-so-called-code-of-practice>.
- "Spamhaus." *Spamhaus*, Accessed on June 6, 2021. <https://www.spamhaus.org/>.
- Spiegel, Evan. "How Snapchat is separating social from media." *Axios*, November 29, 2017. <https://www.axios.com/how-snapchat-is-separating-social-from-media-2513315946.html>.
- Stamos, Alex. "Alex Stamos talks about Facebook's Oversight Board." *Galley by CJR*, 2020. <https://galley.cjr.org/public/conversations/-M74eLMfvkdKplPJrfo4>.
- Starbird, Kate. "Disinformation's spread: bots, trolls and all of us." *Nature*, July 24, 2019. <https://www.nature.com/articles/d41586-019-02235-x>.
- Stieglitz, Stefan; Brachten, Florian; Ross, Björn; and Jung, Anna-Katharina. "Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts." *Australasian Conference on Information Systems* (2017). <https://arxiv.org/ftp/arxiv/papers/1710/1710.04044.pdf>.
- Stolton, Samuel. "EU executive mulls tougher rules for microtargeting of political ads". *Euractiv*, March 03, 2021. <https://www.euractiv.com/section/digital/news/commission-mulls-tougher-rules-for-microtargeting-of-political-ads/>
- Sullivan, Andrew. "Looking the GIFCT in the Mouth." *The Internet Society*, October 11, 2019. <https://www.internetsociety.org/blog/2019/10/looking-the-gifct-in-the-mouth/>.
- Tech Against Terrorism. Accessed on May 11, 2021. <https://www.techagainstterrorism.org/>.
- Thorne, James and Vlachos, Andreas. "Automated Fact Checking: Task Formulations, Methods and Future Directions." *Proceedings of the 27th International Conference on Computational Linguistics* (2018): 3346-3359. <https://www.aclweb.org/anthology/C18-1283>.
- Tiffany, Kaitlyn. "Who Would Volunteer to Fact-Check Twitter?" *The Atlantic*, March 3, 2021. <https://www.theatlantic.com/technology/archive/2021/03/twitters-birdwatch-aims-to-crowdsource-fact-checking/618187/>.
- TikTok. "Community Guidelines." Last updated December 2020. <https://www.tiktok.com/community-guidelines?lang=en>.
- TikTok. *December 2020 Report EU Code of Practice on Disinformation / COVID-19*. TikTok: 2021.
- TikTok. "TikTok Transparency Report 2020 H1." September 22, 2020. <https://www.tiktok.com/safety/resources/transparency-report-2020-1?lang=en>.
- Tiku, Nitasha. "Why Snapchat And Apple Don't Have A Fake News Problem." *Buzzfeed*, December 1, 2016. <https://www.buzzfeednews.com/article/nitashatiku/snapchat-fake-news>.
- Tsagourias, Nicholas. "Electoral Cyber Interference, Self-Determination, and the Principle of NonIntervention in Cyberspace." In *Governing Cyberspace: Behaviour, Power and Diplomacy*, edited by Dennis Broeders, and Bibi van den Berg, 45–63. London: Rowman & Littlefield, 2020.

- Tschiatschek, Sebastian; Singla, Adish; Rodriguez, Manuel Gomez; Merchant, Arpit; and Krause, Andreas. "Fake News Detection in Social Networks via Crowd Signals." *Companion Proceedings of the The Web Conference 2018* (April 2018). <https://www.microsoft.com/en-us/research/publication/fake-news-detection-social-networks-via-crowd-signals/>.
- No Author. "Turning a blind eye to bots to protect ad revenue? Think again." *What's New in Publishing*, 2019. <https://whatsnewinpublishing.com/turning-a-blind-eye-to-bots-to-protect-ad-revenue-think-again/>.
- Twitter. "Synthetic and manipulated media policy." *Twitter Help Center*, accessed on May 11, 2021. <https://help.twitter.com/en/rules-and-policies/manipulated-media>.
- Twitter. *Twitter Progress Report: Code of Practice on Disinformation*. Twitter: 2020.
- Twitter. *Twitter Report: Staying safe and informed on Twitter during COVID-19*. Twitter: 2021.
- Twitter. "Twitter Transparency Center." Accessed May 11, 2021. <https://transparency.twitter.com/>.
- Twitter. "Updates on Covid-19 in the Netherlands." Accessed May 12, 2021. <https://twitter.com/i/events/1244645077797851137>.
- Twitter Help Center. "About Twitter's APIs." *Twitter*, accessed May 12, 2021. <https://help.twitter.com/en/rules-and-policies/twitter-api>.
- Twitter Developer. "Academic Research." Accessed on May 12, 2021. <https://developer.twitter.com/en/solutions/academic-research>.
- Twitter Help Center. "Civic Integrity Policy." *Twitter*, January 2021. <https://help.twitter.com/es/rules-and-policies/election-integrity-policy>.
- Twitter Help Centre. "Twitter Moments guidelines and principles." *Twitter*, accessed May 12, 2021. <https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles>.
- Twitter Help Center. "The Twitter Rules." *Twitter*, accessed on May 12, 2021. <https://help.twitter.com/en/rules-and-policies/twitter-rules>.
- Twitter Safety. "Disclosing new data to our archive of information operations." *Twitter*, September 20, 2019. [https://blog.twitter.com/en\\_us/topics/company/2019/info-ops-disclosure-data-september-2019.html](https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html).
- Twitter Safety. "Strengthening our approach to deliberate attempts to mislead voters." *Twitter*, April 24, 2019. [https://blog.twitter.com/en\\_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html](https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html).
- u/worstnerd (Reddit Admin: Safety). "Misinformation and COVID-19: What Reddit is Doing." *Reddit*, 2020. [https://www.reddit.com/r/ModSupport/comments/g21ub7/misinformation\\_and\\_covid19\\_what\\_reddit\\_is\\_doing/](https://www.reddit.com/r/ModSupport/comments/g21ub7/misinformation_and_covid19_what_reddit_is_doing/).
- United Nations. "Charter of the United Nations." August 10, 2015. <https://www.un.org/en/charter-united-nations/>.
- United Nations General Assembly. "Letter dated 9 January 2015 from the Permanent Representatives of China, Kazakhstan, Kyrgyzstan, the Russian Federation, Tajikistan and Uzbekistan to the United Nations addressed to the Secretary-General." January 13, 2015. [https://eucyberdirect.eu/content\\_knowledge\\_hu/2015-sco-international-code-of-conduct-for-state-behaviour-in-information-security/](https://eucyberdirect.eu/content_knowledge_hu/2015-sco-international-code-of-conduct-for-state-behaviour-in-information-security/).
- United Nations General Assembly. "United Nations General Assembly Resolution 53/70." United Nations: January 4, 1999. <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N99/760/03/PDF/N9976003.pdf?OpenElement>.
- United Nations Group of Governmental Experts. "Report of the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/68/98." United Nations: June 24, 2013. <https://undocs.org/A/68/98>.
- United Nations Group of Governmental Experts. "United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security A/70/174." United Nations: July 25, 2015. <https://undocs.org/A/70/174>.

- United Nations Group of Governmental Experts. "Report of the Group of Governmental Experts on Advancing responsible State behaviour in cyberspace in the context of international security". United Nations: May 28, 2021 (advance copy). <https://front.un-arm.org/wp-content/uploads/2021/06/final-report-2019-2021-gge-1-advance-copy.pdf>
- United Nations Security Council. "Resolution 2129." United Nations: December 17, 2013. UN doc: S/RES/2129(2013)
- United Nations Security Council. "Resolution 2354". United Nations: May 24, 2017. UN doc: S/RES/2354(2017)
- United States Department of Justice. "Case 1:18-cr-00032-DLF." United States District Court of for the District of Columbia, February 16, 2018. <https://www.justice.gov/file/1080281/download>.
- Vijayan, Jaikumar. "Google Introduces Fact Check Label on News Stories, Search Results." *eWeek*, April 7, 2017. <https://www.eweek.com/cloud/google-introduces-fact-check-label-on-news-stories-search-results/>.
- Villasenor, John. "How to deal with AI-enabled disinformation." *Brookings*, November 23, 2020. <https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation/>.
- No Author. "Voters Don't Trust Media Fact-Checking." *Rasmussen Reports*, September 30, 2016. [https://www.rasmussenreports.com/public\\_content/politics/general\\_politics/september\\_2016/voters\\_don\\_t\\_trust\\_media\\_fact\\_checking](https://www.rasmussenreports.com/public_content/politics/general_politics/september_2016/voters_don_t_trust_media_fact_checking).
- Wales, Jimmy. "Wikipedia's strength is in collaboration – as we've proved over 15 years." *The Guardian*, January 15, 2016. <https://www.theguardian.com/commentisfree/2016/jan/15/wikipedia-israel-palestine-15-years-encyclopedia>.
- Walker, Mason and Gottfried, Jeffery. "Republicans far more likely than Democrats to say fact-checkers tend to favor one side." *Pew Research Center*, June 27, 2019. <https://www.pewresearch.org/fact-tank/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/>.
- Walter, Nathan; Cohen, Jonathan; and Morag, Yasmin. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37, no. 3 (2020): 350-375. <https://doi.org/10.1080/10584609.2019.1668894>.
- Wardle, Claire. "Understanding Information disorder." *First Draft*, September 22, 2020. <https://firstdraft-news.org/long-form-article/understanding-information-disorder/>.
- Wardle, Claire and Derakhshan, Hossein. *Information Disorder: Toward an interdisciplinary framework for research and policymaking*. Strasbourg: Council of Europe, 2017. <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>.
- Wikipedia. "Net neutrality." Accessed May 6, 2021. [https://en.wikipedia.org/wiki/Net\\_neutrality#:~:text=Network%20neutrality%2C%20most%20commonly%20called,source%20address%2C%20destination%20address%2C%20or](https://en.wikipedia.org/wiki/Net_neutrality#:~:text=Network%20neutrality%2C%20most%20commonly%20called,source%20address%2C%20destination%20address%2C%20or).
- Williams, Demetrius. "How Social Media Fact-checking is Inconsistent Across Languages". *TranslateMedia*, June 01, 2021. <https://www.translatemedia.com/translation-blog/how-social-media-fact-checking-is-inconsistent-across-languages/>
- Wong, Queenie. "Twitter wants to make it easier for researchers to analyze tweets." *Cnet*, January 26, 2021. <https://www.cnet.com/news/twitter-wants-to-make-it-easier-for-researchers-to-analyze-tweets/>.
- World Federation of Advertisers. "Global Alliance for Responsible Media." Accessed May 11, 2021. <https://wfanet.org/leadership/garm/about-garm>.
- Yaraghi, Niam. "Twitter's ban on political advertisements hurts our democracy." *Brookings*, January 8, 2020. <https://www.brookings.edu/blog/techtank/2020/01/08/twitters-ban-on-political-advertisements-hurts-our-democracy/>.
- York, Jillian C. "Terrorists on Twitter: Attempts to silence ISIS online could backfire." *Slate*, June 25, 2014. <https://slate.com/technology/2014/06/isis-twitter-suspended-how-attempts-to-silence-terrorists-online-could-backfire.html>.

- YouTube. "Community Guidelines." Accessed May 12, 2021. <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#community-guidelines>.
- YouTube. "Expanding fact checks on YouTube to the United States." April 28, 2020. <https://blog.youtube/news-and-events/expanding-fact-checks-on-youtube-to-united-states>.
- YouTube Help. "Report inappropriate content." *YouTube*, accessed on May 11, 2021. <https://support.google.com/youtube/answer/2802027#zippy=>.
- YouTube Help. "See fact checks in YouTube search results." *YouTube*, accessed May 12, 2021. <https://support.google.com/youtube/answer/9229632?hl=en>.
- Zalan, Eszter. "EU Commission plans sanctions on disinformation." *EUobserver*, December 4, 2020. <https://euobserver.com/political/150279>.
- Zhang, Dongyan and Wang, Shuo. "Optimization of Traditional Snort Intrusion Detection System." *IOP Conference Series: Materials Science and Engineering* 569, no. 4 (2019). <https://doi.org/10.1088/1757-899X/569/4/042041>.
- Zingales, Nicolo. "Virtues and Perils of Anonymity: Should Intermediaries Bear the Burden?" *Journal of Intellectual Property, Information Technology and E-Commerce Law* 155 (2014). <https://www.jipitec.eu/issues/jipitec-5-3-2014/4091>.
- Zuckerberg, Mark. "A Blueprint for Content Governance and Enforcement." *Facebook*, May 5, 2021. <https://www.facebook.com/notes/751449002072082/>.
- Zuckerberg, Mark. "Facebook's commitment to the Oversight Board." *Facebook*, 2019. <https://about.fb.com/wp-content/uploads/2019/09/letter-from-mark-zuckerberg-on-oversight-board-charter.pdf>.



The Hague Centre  
for Strategic Studies

**HCSS**

Lange Voorhout 1  
2514 EA Hague

**Follow us on social media:**

@hcssnl

**The Hague Centre for Strategic Studies**

Email: [info@hcss.nl](mailto:info@hcss.nl)

Website: [www.hcss.nl](http://www.hcss.nl)