WORKING GROUP ON INFODEMICS

POLICY FRAMEWORK

NOVEMBER 2020

Forum on Information & Democracy

CONTENTS

HOW TO ADDRESS
THE INFORMATION CHAOS

By Christophe Deloire, Chair of the Forum on Information and Democracy

With the creation of ***the International Commission on Information and Democracy*** *in September 2018, we undertook to define the principles of the global information and communication space, 'this common good of humankind', in order to impose democratic safeguards.*

A few days after its publication, in November 2018, ***the Declaration on Information and Democracy was supported by the secretary-general of the United Nations, António Guterres,*** *the director-general of UNESCO, Audrey Azoulay, and twelve heads of state and government.*

The French president, Emmanuel Macron, presented the initiative at the G7 summit in Biarritz, before it gave rise to the launch of the Partnership on Information and Democracy, at a meeting of the Alliance for Multilateralism on the sidelines of the United Nations General Assembly in 2019.

Created at the end of 2019 by eleven organizations, *research centers and think tanks, the Forum aims at implementing the Partnership, now signed by 38 countries. It launched its inaugural working group on infodemics in June 2020.*

At a meeting of the Alliance for Multilateralism, bringing together 50 foreign ministers at the instigation of the French and German ministers Jean-Yves Le Drian and Heiko Maas, many representatives welcomed the creation of this working group and gave assurance of waiting for its recommendations.

This report is the fruit of the work of a Steering commitee, co-chaired by Maria Ressa and Marietje Schaake, dozens of researchers and lawyers all over the planet, whose efforts have been synthesized by a team of rapporteurs. I would like to pay particular tribute to all of them.

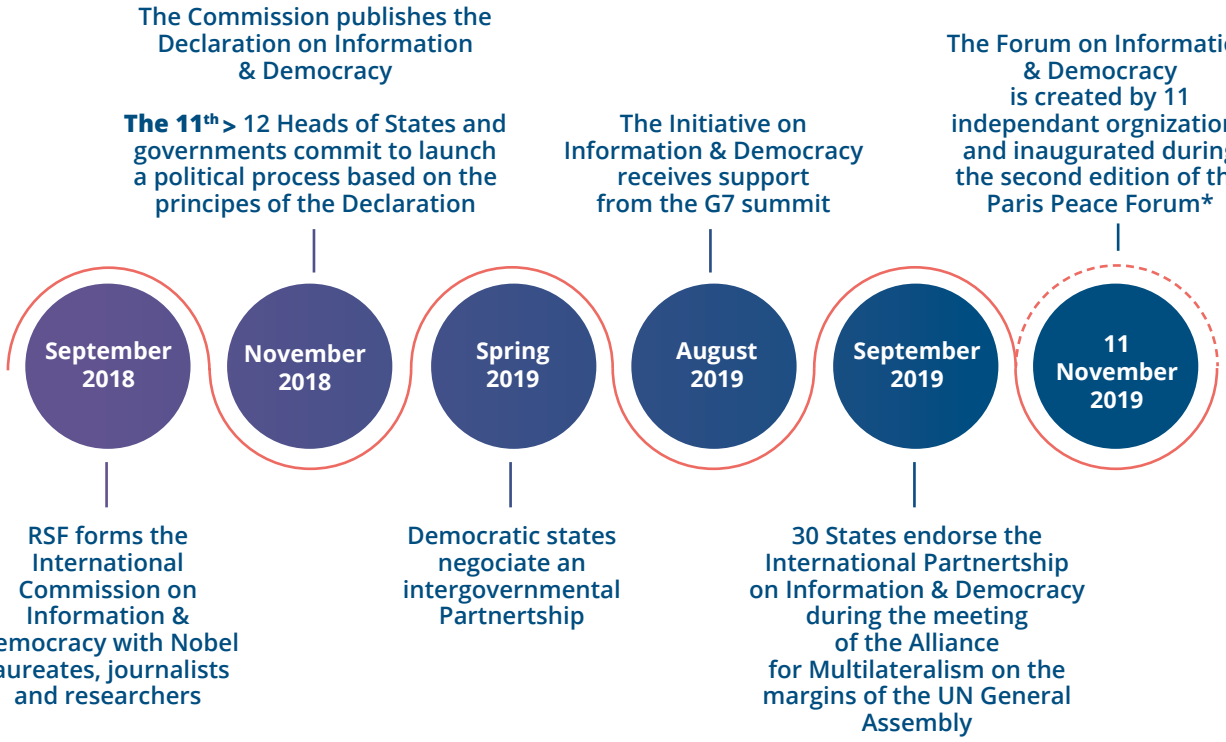
The Information and Democracy Initiative demonstrates that a structural ***solution is possible to end the informational chaos that poses a vital threat to democracies.*** *The exercise of human rights, presupposes that democratic systems impose rules on the entities that create the standards and the architectures of choice in the digital space.*

This initiative demonstrates ***a capacity for reinventing multilateralism, with an innovative articulation between States and civil society.*** *Initiated by Reporters Without Borders (RSF), this process resulted in an intergovernmental text.*

The signatory states of the Information and Democracy Partnership now represent a coalition that can exert influence to implement a democratic vision in the digital space.

The Forum on Information and Democracy has complete independence from States. However, its work is intended to be the raw material for regulation. The Forum thus has a major role to play in facing the democratic emergency.

ABOUT THE INITIATIVE
ON INFORMATION AND DEMOCRACY



*CIGI, CIVICUS, the Digital Rights Foundation, Free Press Unlimited, the Human Rights Centre at UC Berkeley School of Law, the institute for Strategic Dialogue, OBSERVACOM, the Open Government Partnership, the Peace Research Institute Oslo, Reporters without Borders (RSF), Research ICT Africa

THE PARTNERSHIP ON INFORMATION AND DEMOCRACY[1]

THE SIGNATORY STATES OF THE PARTNERSHIP ON INFORMATION & DEMOCRACY[2] :

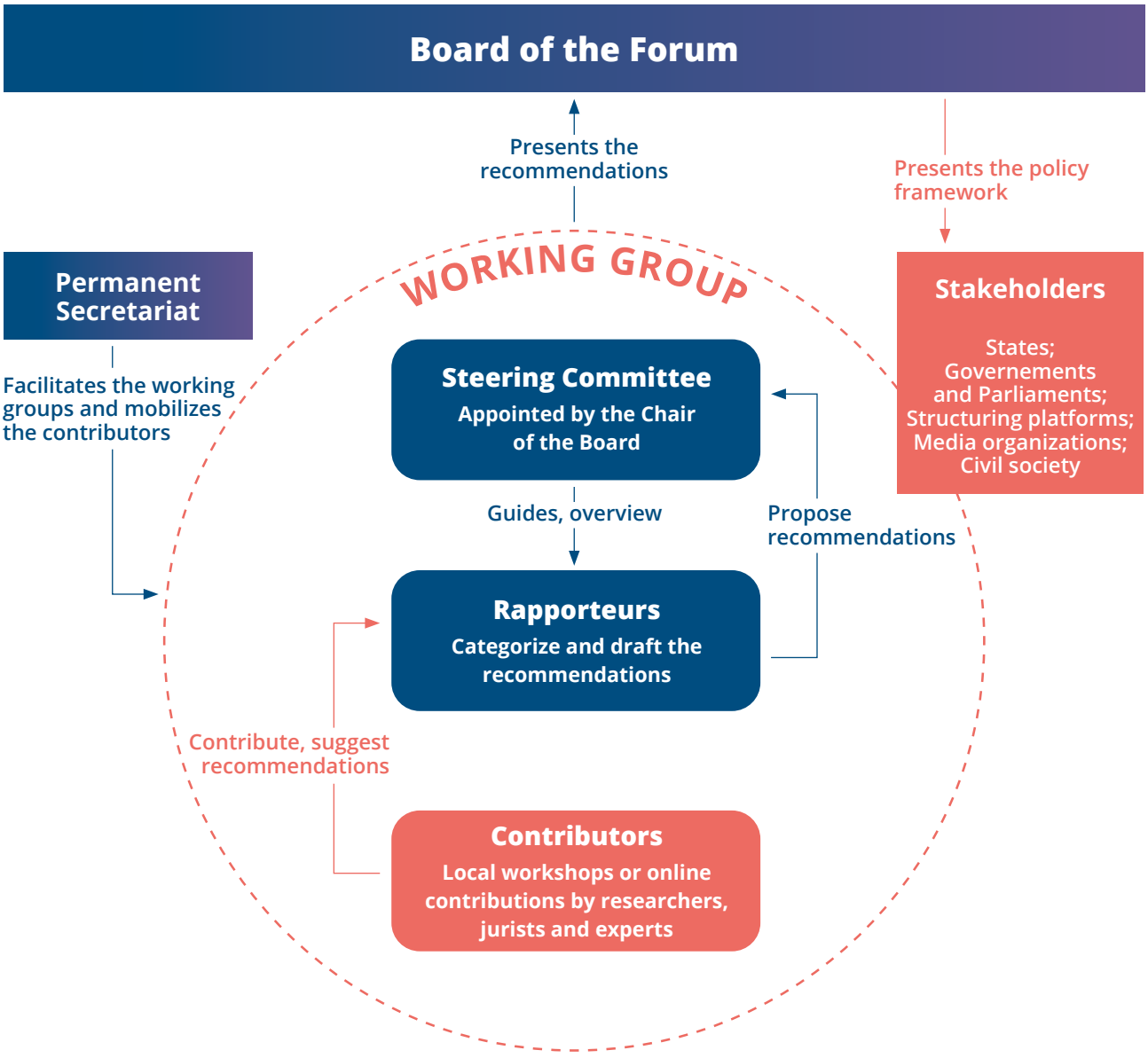
ANDORRA	ARMENIA	AUSTRALIA	AUSTRIA	BENIN
BULGARIA	CANADA	CHILE	COSTA RICA	CROATIA
CZECH REPUBLIC	DENMARK	ESTONIA	FINLAND	FRANCE
GERMANY	GREECE	INDIA	ITALY	LATVIA
LEBANON	LIECHTENSTEIN	LITHUANIA	LUXEMBOURG	MALTA
MONTENEGRO	NETHERLANDS	NORWAY	NORWAY	SENEGAL
SLOVENIA	SOUTH AFRICA	SOUTH KOREA	SPAIN	SWEDEN
SWITZERLAND	TUNISIA	UNITED KINGDOM		

1 See the Partnership text on page 119.
2 List of countries that have signed the International Partnership on Information & Democracy as of 1 November 2020.

THE FORUM ON INFORMATION AND DEMOCRACY
MANDATE

Evaluating	Investigating	Making	Facilitating	Supporting
the means, norms and architectures of the global information and communication space	corporate entities' behaviour as regards adherence to the principles of the Declaration and Patnership	recommendations to the different stakeholders that shape this space on how the norms should evolve	the emergence of regulatory and self-regulatory responses by and for the different stakeholders	the social function of journalism through innovative responses and recommendations

HOW WE WORK



WORKING GROUP ON INFODEMICS

POLICY FRAMEWORK

Maria Ressa & Marietje Schaake, *Steering Committee co-chairs*
Delphine Halgand-Mishra, *lead rapporteur*
Iris de Villars & Jenny Domino, *rapporteurs*
Dan Shefet, *legal advisor*

STEERING COMMITTEE

Maria Ressa, co-chair
Journalist, CEO of the investigation website Rappler in the Philippines. *Time* magazine Person of the Year in 2018. Member of the Commission on Information and Democracy

Marietje Schaake, co-chair
Former member of the European Parliament (2009 – 2019). Currently international policy director of the Stanford Cyber Policy Center and president of the Cyber Peace Institute.

Sinan Aral
David Austin Professor of Management, Marketing, IT and Data Science at MIT, director of the MIT Initiative on the Digital Economy (IDE) and a founding partner at Manifest Capital.

Julia Cagé
Author of bestselling books about democracy and media. Professor of economics at Sciences Po, and co-director of LIEPP 'Evaluation of Democracy' Research Group.

Ronald Deibert
Director of the Citizen Lab at the Munk School of Global Affairs & Public Policy. Co-founder and principal investigator of the OpenNet Initiative and Information Warfare Monitor projects.

Camille François
Chief innovation officer at Graphika, leading the company's work to detect and mitigate disinformation and media manipulation. Previously the principal researcher at Jigsaw.

David Kaye
Former United Nations special rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. Clinical professor of law (University of California).

Roukaya Kasenally
CEO of the African Media Initiative. Professor in media and political systems (University of Mauritius). Chair of the board of the Electoral Institute for Sustainable Democracy in Africa.

Edison Lanza
Lawyer, former special rapporteur for Freedom of Expression of the Inter- American Commission on Human Rights. Founder of several NGOs which defend freedom of expression.

Roger McNamee
Author of Zucked: Waking up the Facebook Catastrophe, and tech venture capitalist. Head of the T. Rowe Price Science and Technology Fund. Former investor in Facebook.

Jun Murai
Co-director, Keio University Cyber Civilization Research Center. Founder of the WIDE project. Known as the 'father of the internet in Japan'.

Peter Pomerantsev
Visiting senior fellow at the Institute of Global Affairs at the London School of Economics. Senior fellow at the Agora Institute at Johns Hopkins University.

Julie Posetti
Global director of research at the International Center for Journalists. Previously, lead of the Journalism Innovation Project at the Reuters Institute for the Study of Journalism.

Anya Schiffrin
Director of the Technology, Media, and Communications specialization at Columbia University's School of International and Public Affairs.

Vivian Schiller
Executive director of Aspen Digital program at the Aspen Institute. Former president and CEO of NPR. Board member of Reporters Without Borders (RSF), USA.

Wolfgang Schulz
Director of the Humboldt Institute for Internet and Society. Lecturer in the field of information and communication at the law faculty of the University of Hamburg.

Christopher Wylie
Data scientist and the whistleblower who reported Cambridge Analytica and Facebook. Listed in TIME 100 Most Influential People in the World, and Forbes 30 Under 30.

The report of this inaugural working group reflects views expressed during the rapporteurs' team discussions with the Steering Committee, more than 60 experts, and in written contributions received from experts and organizations most engaged in the field. The team of rapporteurs did not seek unanimity on every conclusion or recommendation, recognizing that diverse perspectives could not always be reconciled. This report should not be understood to be the result of a formal negotiation validated by the Steering Committee members, but as the rapporteur team's best efforts to offer a path forward.

FOREWORD
BY MARIA RESSA

Co-chair of the steering committee of the working group on infodemics

I know first-hand why and how democracy is dying, and why journalists are constantly under attack. Part of the reason for that is because of my front-row seat in the Philippines: as a target, I see the evolution of the online attacks as well as the weaponization of the law; as the editorial head of a news group, I'm faced with near daily decisions challenging our profession; and, as Rappler's business and technology head, I'm watching our business model being destroyed by technology.

Social media, once an enabler, is now the destroyer, building division—'us against them' thinking—into the design of their platforms. It's not a coincidence that divisive leaders perform best on social media.

Facebook is now the world's largest distributor of news. Except there's a catch: lies laced with anger and hate spread faster and further than the boring facts of news. They create a bandwagon effect of artificial consensus—for the lie.

You repeat a lie a million times, it becomes a fact. Without facts, you can't have truth. Without truth, you can't have trust. Without these, democracy as we know it is dead.

In 2016, I warned that what is happening in the Philippines is coming soon to a democracy near you: the brutal killings by law enforcers enabled by the exponential hate pumping through social media targeting journalists, human rights activists, and opposition politicians.

Our dystopian present is your future, and it has come to pass.

All around the world, populist digital authoritarians use this scorched-earth policy to get elected, then they use the formal powers of their posts—the tools of democracy—to cave institutions in from within.

It's time to end the whack-a-mole approach of the technology platforms to fix what they have broken.

This is why this initiative of the Forum on Information & Democracy is personally satisfying: we found experts obsessed with finding structural solutions to fix our information dystopia. Working with a team of rapporteurs led by Delphine Halgand-Mishra, we submit these ideas.

The 38-nation intergovernmental approach of the Partnership on Information & Democracy reminds me of how 75 years ago, after the end of World War II, the world got together to find multilateral, international solutions to prevent humanity from destroying itself, including NATO, Bretton Woods, and the Universal Declaration of Human Rights.

Working on this with my co-chair, Marietje Schaake, and the stellar steering committee gives me hope: even as we stand on the rubble of the world that was, we can build better—a world more equal, more sustainable, more compassionate.

Join us.

FOREWORD
BY MARIETJE SCHAAKE

Co-chair of the steering committee of the working group on infodemics

For too long, technology was put first when discussing democracy and digitization. Many political leaders and business executives, while based in democratic countries, hoped and promised that without explicit safeguards, technology would liberalize societies and emancipate individuals. That was an illusion. In non-democratic countries, people were clearly confronted with the harms of surveillance and repression, often helped by Western-made systems. Important lessons were missed by hoping, not regulating, for democratic principles to survive.

The past years have offered a wake-up call for those who needed it. Foreign interferers in the US Presidential Election of 2016, and in many elections around the world, used social media platforms as their vehicles of choice. Commercially governed ecosystems determine the information flows for billions of people. This connectivity leads to new vulnerabilities. Without explicit and enforceable safeguards, the technologies promised to advance democracy will prove to be the ones that undermine it.

It is now vital that democracy is made more resilient. The Forum on Information & Democracy proposes a number of policy steps to democratic governments and their supporters. Transparency and accountability need to be shored up and content moderation should be done according to democratic mandates and oversight. The impact of new platforms where disinformation can go viral, such as private messenger services, needs to be understood.

Through a global democratic coalition, a meaningful alternative should be offered instead of the two dominant models of technology governance: the privatized and the authoritarian. Through the intergovernmental Partnership on Information & Democracy, democratic leaders recognize the information and communication space as a 'public good'. Now they have to implement their commitments in policies on the national and international level. Our recommendations are designed to shape and support their policy agenda. Civil society organizations and a broad group of like-minded individuals should make clear that democracy is worth fighting for.

Now is the time to implement a positive agenda for change!

TWELVE MAIN RECOMMENDATIONS

PUBLIC REGULATION IS NEEDED TO IMPOSE TRANSPARENCY REQUIREMENTS ON ONLINE SERVICE PROVIDERS.

1. Transparency requirements should relate to all platforms' core functions in the public information ecosystem: content moderation, content ranking, content targeting, and social influence building.

2. Regulators in charge of enforcing transparency requirements should have strong democratic oversight and audit processes.

3. Sanctions for non-compliance could include large fines, mandatory publicity in the form of banners, liability of the CEO, and administrative sanctions such as closing access to a country's market.

A NEW MODEL OF META-REGULATION WITH REGARDS TO CONTENT MODERATION IS REQUIRED.

4. Platforms should follow a set of Human Rights Principles for Content Moderation based on international human rights law: legality, necessity and proportionality, legitimacy, equality and non discrimination.

5. Platforms should assume the same kinds of obligation in terms of pluralism that broadcasters have in the different jurisdictions where they operate. An example would be the voluntary fairness doctrine.

6. Platforms should expand the number of moderators and spend a minimal percentage of their income to improve quality of content review, and particularly, in at-risk countries.

NEW APPROACHES TO THE DESIGN OF PLATFORMS HAVE TO BE INITIATED.

7. Safety and quality standards of digital architecture and software engineering should be enforced by a Digital Standards Enforcement Agency. The Forum on Information and Democracy could launch a feasibility study on how such an agency would operate.

8. Conflicts of interests of platforms should be prohibited, in order to avoid the information and communication space being governed or influenced by commercial, political or any other interests.

9. A co-regulatory framework for the promotion of public interest journalistic contents should be defined, based on self-regulatory standards such as the Journalism Trust Initiative; friction to slow down the spread of potentially harmful viral content should be added.

SAFEGUARDS SHOULD BE ESTABLISHED IN CLOSED MESSAGING SERVICES WHEN THEY ENTER INTO A PUBLIC SPACE LOGIC.

10. Measures that limit the virality of misleading content should be implemented through limitations of some functionalities; opt-in features to receive group messages, and measures to combat bulk messaging and automated behavior.

11. Online service providers should be required to better inform users regarding the origin of the messages they receive, especially by labelling those which have been forwarded.

12. Notification mechanisms of illegal content by users, and appeal mechanisms for users that were banned from services should be reinforced.

DEFINITIONS

Infodemic: Overabundance of information—some accurate and some not—occurring during an epidemic. This makes it hard for people to find trustworthy sources and reliable guidance when they need it.[3]

Disinformation: Information that is false and deliberately created to harm a person, social group, organization or country.[4]

Misinformation: Information that is false but not created with the intention of causing harm.[5]

Online service providers: Entities that help structure the information and communication space by creating the technical means, architecture, and standards for information and communication.[6]
This includes **Digital platforms ("platforms"),** which are defined as: 'online sites and services that:

(a) host, organize, and circulate users' shared content or social interactions,
(b) without having produced or commissioned (the bulk of) that content,
(c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit.

Platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement.' [7]
In this report, "platforms" and 'online service providers' are used interchangeably.

State regulation: Regulation enforced by governments.

Self-regulation: Regulation exercised by platforms.[8]

Co-regulation: A system in which the general guidelines and expected results of platform policies are defined in a legal instrument, with input from multiple sectors, which must be applied directly by platforms taking into consideration local and regional context and in line with human rights principles. An appropriate body, with guarantees of independence and autonomy, should oversee the companies' application of these standards.[9] Co-regulation should include civil society and could potentially exclude governments.

Meta-regulation: A set of baseline principles.

---

3 As defined by the World Health Organization:
1st WHO Infedemiology Conference. (2020). Retrieved from https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference (Accessed on 22 October 2020).
4 As defined by UNESCO:
Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training. (2020).
Retrieved from https://en.unesco.org/fightfakenews (Accessed on 21 October 2020).
5 As defined by UNESCO, *op. cit.*
6 As defined in the Partnership on Information and Democracy:
International Partnership on Information & Democracy. (2019).
Retrieved from https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed on 21 October 2020).
7 Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. p. 18-21.
Retrieved from: https://yalebooks.yale.edu/book/9780300173130/custodians-internet
8 As summarized in:
Pírková, E. & Pallero, J. (2020). 26 Recommendations on Content Governance : a Guide for Lawmakers, Regulators, and Company Policy Makers. *Access Now*. Retrieved from https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (Accessed on 21 October 2020).
9 As defined in:
Observacom et.al. (July 2020). *Standards for the Democratic Regulation of Large Content Platforms to Ensure Freedom of Expression Online and an Open and Free Internet*. Contribution to this working group.

Chapter 1:
Transparency of Platforms

Access to the qualitative and quantitative data of the leading digital platforms and access to their algorithms is a prerequisite for evaluating them. Transparency requirements must, therefore, be imposed on such platforms in order to be able to determine whether they are respecting their responsibilities in these areas and, in general, with regard to their business models and algorithmic choices.

Contents

INTRODUCTION

WHY TRANSPARENCY?

It is time to enforce real and legally binding transparency of the online service providers that structure the global information and communication space.[10]

It is time to move from self-regulation to public regulation for transparency requirements of digital platforms[11], while making sure to protect freedom of expression and not stifle innovation.

Public regulation related directly to content moderation (a take-down of content approach) is far more dangerous and could lead to building a censorship framework that might harm freedom of expression and other rights and freedoms. Transparency should be adopted according to rule-of-law standards as understood for democratic societies, and transparency must be an obligation of governments as well. Government demands can be as problematic as company policies in some jurisdictions.

A new approach to legally binding transparency is needed to solve many online content moderation and disinformation issues. **This is the first step to better oversight, greater accountability, and to regaining trust[12] between platforms, governments and the public.** It is the first step towards strong evidence-based policies and potentially further regulation by governments. It is the first step in enabling online service providers to face problems and weaknesses[13] they can no longer hide. A radical new approach to transparency is a first step in empowering citizens.

Legally enforced transparency is not a silver bullet that will fix all the issues, but is a necessary condition to develop a more balanced equilibrium of power between the private platforms and democratic societies. The behavior of platforms in fighting disinformation will continue to evolve as online risks and harms continue to evolve; the regulatory framework should not aim at dictating their behavior, but rather at creating the necessary condition for meaningful and open policy dialogues to take place. Legally enforced transparency is a natural corollary to the power that platforms hold over our information ecosystem. 'The legislative frameworks for intellectual property or trade secrets should not preclude such transparency, nor should States or private parties seek to exploit them for this purpose. Transparency levels should be as high as possible and proportionate to the severity of adverse human rights impacts,' notes the Council of Europe in its recommendations to member states.[14]

This chapter reviews the transparency requirements of digital platforms related to their core functions in the public information ecosystem of content moderation, content ranking, content targeting, and social influence building; and proposes transparency regulation models in which a sound audit process open to researchers is key.

10 As stated in the International Partnership on Information and Democracy:
International Partnership on Information & Democracy. (2019). Retrieved from https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed on 21 October 2020).

11 See Key Findings of 2019 Ranking Digital Rights Corporate Accountability Index: Companies still do not adequately inform people about all the ways user information is collected and shared, with whom, and why. As companies struggle to curb extremism, hate speech, and disinformation, most lacked transparency about how they police content or respond to government demands. Retrieved from: https://rankingdigitalrights.org/index2019/ (Accessed on November 2, 2020)

12 Based on discussion with Jun Murai. It is interesting to follow the work currently led by Japan to build trust to design the future structure of data governance, looking at decentralized ID and traceability.

13 Silverman, C., R. Mac, & P. Dixit. (2020). A Whistleblower Says Facebook Ignored Global Political Manipulation. *Buzzfeed News*. Retrieved from https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo (Accessed on 18 October 2020).

14 Recommendation CM/Rec(2020)1 of the Committee of Ministers to member states on the human rights impacts of algorithmic systems. (2020).
Retrieved from https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154 (Accessed on 21 September 2020).

> *None of the information made public by the platforms concerning their self-regulatory actions can be corroborated by objective facts.*[15]

Benoit Loutrel's report, Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France With a European Vision.

> *The lack of transparency is itself a form of censorship, as it means that a citizen simply can't engage with the information forces around them as an equal. Such wise regulation, however, won't be a cure-all—it will just even out the playing field so that those of us who want to save deliberative democracy can start to compete with the forces that seek to sow mistrust and extreme polarization.' Technology researcher.*

Peter Pomerantsev.[16]

TRANSPARENCY OF WHOM?

This new era of transparency requirements should apply to online service providers that structure the global information and communication space,[17] such as social media platforms like Facebook, YouTube, Twitter, Instagram, TikTok, Reddit, Weibo, Baidu Tieba, Quora, iQIYI, QZone, and VK,[18] as well as search engines that have their own moderation practices, such as Google. Specific transparency requirements of private messaging platforms (WhatsApp, Telegram, etc.) are addressed in Chapter 4. These should be included in the same transparency regulation models.

This new era of transparency requirements should take into account the diversity of the platforms and, at least initially, concentrate on those with the most influence over our societies, according to the number of users relative to the size and scale of the market.[19] This is where the failure to provide appropriate content moderation and the corresponding failure to adequately protect freedom of expression create the greatest harm.[20]

As stated in the International Partnership on Information & Democracy: When creating technical means, architectures that shape choices and norms for communication, entities that contribute to the structure of the information and communication space shall respect the principles and guarantees that nourish and underpin the democratic nature of this space. They have to be held accountable in accordance with and in proportion to the impact of their contribution or participation.[21]

For smaller companies, a sliding scale could be implemented with basic requirements expected of all service providers; a second level of disclosure for midsize companies, and full compliance for the largest ones. Care should be taken that the regulatory system does not create an insurmountable entry barrier for mid-sized market players or new entrants, and, consequently, unduly reinforce the power of hegemonic actors by creating regulatory barriers.

---

15 *Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France With a European Vision*. (2019). Retrieved from https://minefi.hosting.augure.com/Augure_Minefi/r/ContenuEnLigne/Download?id=AE5B7ED5-2385-4749-9CE8-E4E1B36873E4&filename=Mission%20Re%CC%81gulation%20des%20re%CC%81seaux%20sociaux%20-ENG.pdf (Accessed on 20 September 2020). p. 12.

16 Pomerantsev, P. (2019). The Death of the Neutral Public Sphere*. The American Interest.* Retrieved from https://www.the-american-interest.com/2019/09/18/the-death-of-the-neutral-public-sphere/ (Accessed on 15 August 2020).

17 International Partnership for Information and Democracy. *op. cit.*

18 Thirteen of the top 50 online content-sharing services are Chinese, and none of them issues TVEC transparency reports. Documents de travail de l'OCDE sur l'économie numérique, n° 296. Current Approaches to Terrorist and Violent Extremist Content Among the Global Top 50 Online Content-Sharing Services. (2020). Retrieved from https://www.oecd-ilibrary.org/docserver/68058b95-en.pdf?expires=1603052099&id=id&accname=guest&checksum=48EEE4D7F91F006EFB0F497905D7A283 (Accessed on 10 October 2020).

19 *Creating a French Framework*. *op. cit.*, p. 12.

20 MacCarthy, M. (2020). *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*. Annenberg Public Policy Center. Retrieved from https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Transparency_TWG_MacCarthy_Feb_2020.pdf (Accessed on 18 August 2020).

21 International Partnership for Information and Democracy, *op. cit*.

Governments should also publish detailed transparency reports, made publicly available, on all content-related requests issued to the online service providers that structure the global information and communication space. This is recommended both by the Manilla Principles[22] defined by seven non-governmental organizations, and the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression in his 2018 report. These questions are discussed in section 2.1 of this chapter.[23]

1. THE PERIMETERS OF TRANSPARENCY

'Internet companies make calls every day that influence who has the ability to speak and what content can be shared on their platform. Problems arise when people do not understand the decisions that are being made or feel powerless when those decisions impact their own speech, behavior, or experience,' even Monika Bickert, Vice-President Content Policy at Facebook, acknowledges.[24]

According to a recommendation of the Council of Europe on the roles and responsibilities of Internet intermediaries, a new approach should require platforms to **disclose information and data in a way that meaningfully represents their interferences with the exercise of rights and freedoms in the digital environment.**[25]

The focus of the public debate has mainly been on the «content moderation» function of digital platforms. However, the power that online service providers wield is as important in their three other core functions, as this also has an impact on the spread of disinformation: 'ranking' (how they organize, rank and present the user generated content at scale,) 'targeting' (how they push unsolicited contents for third parties on a commercial basis), and 'socializing' (how they influence the development of each user's social network, for example in suggesting 'don't you want to be connected to these individuals…').

The transparency requirements of platforms should throw light on all four key functions, not just content moderation.

1.1. GENERAL PRINCIPLES

RECOMMENDATIONS
TO **STATES**

> Impose a general principle of transparency by law:

- Transparency obligations should be imposed for every public or private sector entity in proportion to the power or influence it is able to exercise over people or ideas.

22 Manila principles on Intermediary Liability. Retrieved from https://www.manilaprinciples.org/ (Accessed on 19 October 2020).
23 Kaye, D. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.* Retrieved from https://www.undocs.org/A/HRC/38/35 (Accessed on 10 August 2020).
24 Bickert, M. (2020). Charting a Way Forward Online Content Moderation. Facebook. Retrieved from https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf (Accessed on 19 October 2020).
25 Recommendation CM/Rec(2018)2 of the Committee of Ministers to member states on the roles and responsibilities of internet intermediaries. (2018). Retrieved from https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 (Accessed on 21 September 2020).

- Online service providers must be predictable for those over whom they have influence, resistant to any manipulation, and open to inspection.[26]

> Include the following fields in the perimeter of legal obligations for transparency:

- how the platforms implement and follow their own content moderation policies (see part 1.2 and 1.3);
- how the algorithms operate and with what objectives (see part 1.4);
- what content reached the highest number of users per day, per country (see part 1.5);
- what advertisements are seen on the platforms (see part 1.6);
- how users' data is used, and to allow users to obtain information pertaining to them (collected and inferred), opening avenues to real data portability and interoperability (see part 1.7);
- what human rights impact assessments of their policies and products should include (see part 1.8).

> Structure the governance of these transparency requirements through:

- democratic safeguards and transparency requirements from governments themselves (see part 2.1);
- a sound audit mechanism (see part 2.2);
- a three-tier disclosure to users, vetted researchers and regulators (see part 2.3);
- regulation models for Europe and the US as a starting point (see part 2.4).

> Require platforms to make the transparency requirements:

- easily accessible and intelligible for all users;
- granular and machine-readable for vetted researchers[27] and regulators;
- standardized to be comparable across companies,[28] or at least to be audited by vetted researchers and regulators;
- published at least once a year, preferably once a quarter.

What could be perceived as a **transparency paradox** arises, as on the one hand greater access to information and metadata is recommended, while on the other an attempt must be made to prevent future damaging misuse of data, such as in the Cambridge Analytica scandal. **Differential privacy** could allow a safe approach to transparency. 'Differential privacy' describes a promise, made by a data holder, or curator, to a data subject: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.' At their best, differentially private database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views.... Differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population.[29]

26 As stated in the International Declaration on Information and Democracy (see page 122).

27 Civil society representatives are included in the generic term 'vetted researchers' in this chapter.

28 It will be extremely interesting to follow the work of the OECD, which is developing a common framework and set of metrics for voluntary transparency reporting on TVEC.
West, J. (2020). Why We Need More Transparency to Combat Terrorist and Violent Extremist Content Online. *OECD Innovation Blog*. Retrieved from https://oecd-innovation-blog.com/2020/09/15/terrorist-violent-extremist-content-internet-social-media-transparency-tvec/ (Accessed on 1 October 2020).

29 Dwork, C. & A. Roth. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science.* Retrieved from https://privacytools.seas.harvard.edu/files/privacytools/files/the_algorithmic_foundations_of_differential_privacy_0.pdf (Accessed on 15 October 2020). p.15.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Employ a differential privacy approach to fulfill their requirements towards regulators and vetted researchers.**

1.2. TERMS OF SERVICE / COMMUNITY GUIDELINES / INTERNAL POLICIES

Requiring platforms to publish their policies in detail 'informs the public and forces social media companies to act consistently with their own rules'.[30]

RECOMMENDATIONS
TO **STATES**

> **Require platforms to publish their policies:**[31]
- ♦ regarding what user-generated content and behavior is or is not permitted;
- ♦ for rules about content and targeting for advertising;[32]
- ♦ for content moderation, content ranking, content targeting and socializing recommendations;
- ♦ for processing and disclosure of user data.

> **Require platforms to maintain consistency with their publicly announced standards.**[33]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Make these policies available in all languages and dialects of countries where their services are provided. This should also apply to transparency reports and company blogs.**

> Ensure these documents are easily accessible and intelligible to all users, such as organizing thematic areas on a landing page.

> Services for children and teenagers, regardless of their legal limitations to hire such services, need to establish terms and conditions that can be easily understood by this group of people.[34]

30 MacCarthy, M. (2020). A Dispute Resolution Program for Social Media Companies. *Brookings*. Retrieved from https://www.brookings.edu/research/a-dispute-resolution-program-for-social-media-companies/ (Accessed on 20 October 2020).
31 In line with the Legality Principle discussed in Chapter 2.
32 Maréchal, N. & E. Roberts Biddle. (2020). It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge. Key Transparency Recommendations for Content Shaping and Moderation. *Ranking Digital Rights*. Retrieved from https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/key-transparency-recommendations-for-content-shaping-and-moderation/ (Accessed on 2 September 2020).
33 MacCarthy, M. (2020). A Dispute Resolution Program. *op. cit.*
34 Observacom. (July 2020). A Latin American Perspective for Content Moderation Processes that are Compatible with International Human Rights Standards—Contribution to this working group.

> **Explain how internal policies and rules are elaborated, developed, interpreted and implemented.**

> **Explain the procedures, together with the human and technological resources involved.**

> Release their enforcement guidelines/implementation standards along with the policies.[35]

> **Make transparent, accountable and inclusive the process of drafting, amending, and applying the terms of service agreements, community standards, and content-restriction policies.**

> Collaborate and negotiate with consumer associations, human rights advocates and other civil society organizations representing the interests of users and affected parties, as well as with data protection authorities, before adopting and modifying their policies.

> Empower their users to engage in processes of evaluating, reviewing and revising, where appropriate, intermediaries' policies and practices.[36]

> **Notify users when the rules for user-generated content, for advertising content, or for advertisement targeting change, so that users can make an informed decision about whether to continue using the platform.[37]**

1.3. NOTICE OF VIOLATIONS OF TERMS OF SERVICE (TOS) OR LAWS

The Santa Clara Principles[38] established by civil society organizations recommend that companies notify users if their content is taken down and inform them of the possibility of appeal. Transparency would allow monitoring of whether platforms respect these principles in practice.

RECOMMENDATIONS TO **STATES**

> Require platforms to provide granular and standardized data to vetted researchers and regulators on the:
- ♦ number of all notices of violations of TOS and laws received;
- ♦ type of entities that issued them, including private parties, administrative bodies, or courts;
- ♦ **reasons for determining the legality of content, or how it infringes terms of service;**

35 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit*.
36 Recommendation CM/Rec(2018)2 of the Committee of Ministers. *op. cit.*
37 Maréchal, N. & E. Roberts Biddle. (2020). *op. cit.*
38 The Santa Clara Principles on transparency and accountability in content moderation. Retrieved from https://santaclaraprinciples.org/ (Accessed on 19 October 2020). The Santa Clara Principles will be revised in early 2021.

- **ways the legality or illegality of content were determined (per automated system or human review);**
- concrete time frames for notifying the content provider before any action is taken; for filing the counter-notice; the exact time that will pass before the content is restricted, and the time frame for an appeal procedure;
- number of appeals they received and how they were resolved.[39]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Explain this process clearly to all users**, especially to users who file a complaint.

> Send precise information to the complainants on any follow-up procedures, enforcement action and the reasoning behind the action taken.[40]

> Explain to users why their content has been restricted, limited or removed; or why the account or profile has been suspended, blocked or deleted.
- Notifications should include, at least, the specific clause of the community rules that the user allegedly violated.
- Notifications should be detailed enough to allow the user to specifically identify the restricted content and should include information on how the content or account was detected, evaluated and deleted or restricted.
- Users should be provided with clear information on how to appeal the decision.[41]

1.3.a CONTENT TAKEDOWNS, FLAGGED CONTENT, DISABLED ACCOUNTS & CONTENT REMAINING

Numerous recent media articles tend to indicate that the strategy of major platforms 'has never been to manage the problem of dangerous content, but rather to manage the public's perception of the problem'.[42] Providing further information about their content moderation practices could incentivize platforms to improve the implementation of their own rules.

RECOMMENDATIONS
TO **STATES**

> **Require platforms to publish:**
- the total number of users who access the platform in the country by type of access;
- the total number of fake accounts, bots, and botnets detected, removed or suspended;

39 Pírková, E. & J. Pallero. (2020). 26 Recommendations on Content Governance. *op. cit.*
40 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
41 Observacom. (July 2020). A Latin American Perspective. *op. cit.*
42 Marantz, A. (2020). Why Facebook Can't Fix Itself. *The New Yorker.* Retrieved from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself?mc_cid=a2705e31cc&mc_eid=dd9bd17d22 (Accessed on 20 October 2020); and Silverman, C. et al. *A Whistleblower. op. cit.*

- data related to engagement with misinformation, including number of views, shares, reach, and the number of complaints or requests for removal;
- each piece of corrected content, along with the number of users reached by the initial content and by the correction.[43]

> **Require platforms to provide granular and standardized data to researchers and regulators on content takedowns, content flagged, labeled, downranked, delayed, masked with a warning, and accounts disabled:**

- per user notification, state request, or own TOS, broken down by action taken, specifying which rule has been applied;
- per automated decision, specifying the criteria applied, and releasing the accuracy rate;[44]
- per human decision, specifying the criteria applied, and releasing the accuracy rate:[45] platforms should share statistics on human reviewers' inter-rater reliability;
- number of violating posts as a proportion of the total number of posts (as a percentage of viewed posts and percentage of all posts);[46]
- number of views of violating posts as a proportion of all views;
- breakout actioned content measures, by type of action taken (e.g., content taken down, content covered with warning, account disabled);
- actioned content as a proportion of total estimated violating content.

> **Require platforms to provide granular and standardized data on content remaining on the platform after being notified:**

- per requests by governments, specifying the criteria applied;
- in respect of internal policies, specifying which rule has been applied;
- per automated decision, specifying the criteria applied;
- per human decision, specifying the criteria applied.

The Data Transparency Advisory Group (DTAG), which was chartered by Facebook to assess Facebook's Community Standard Enforcement Reports (CSER) and provide recommendations for how to improve its measurement and reporting practices, asked Facebook to release anonymized and aggregated versions of the data on which the metrics in the CSER are based. This would allow external researchers to verify Facebook's representations.[47]

The DTAG report also suggests that platforms should check reviewers' judgments not only against an internal 'correct' interpretation of the internal standard/guidelines/policies, but also against a survey of users' interpretations of these internal standards/guidelines/policies.

These transparency requirements could open the way to further regulatory paths, recommended by Facebook itself: 'Governments could also consider requiring companies to hit specific performance targets, such as decreasing the prevalence of content that violates a site's hate speech policies, or maintaining a specified median response time to user or government reports of policy violations…'. Such regulations could require a certain level of performance in these areas in order to receive certification or to avoid regulatory consequences.[48]

---

43 Recommendations made by Avaaz in their contribution to this Working Group.
44 Bradford, B., et al. (2019). Report Of The Facebook Data Transparency Advisory Group. *The Justice Collaboratory.* Retrieved from https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf (Accessed on 5 September 2020).
45 Ibid.
46 This is because the number of bad posts viewed is affected by recommendation and prioritization algorithms, as well as by the effectiveness of automated removal systems that proactively detect violating content and block it before it is posted. MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
47 Bradford, B., et al. (2019). *op. cit.*
48 Bickert, M. (2020). Charting a Way Forward. *op. cit.*

1.3.b NOTIFICATION OF USERS & REDRESS MECHANISM

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Notify users if their content is removed or downgraded, explaining what specific rule was violated and why that conclusion was reached, as well as specifying if the decision was automated.**

> **Offer an effective process of redress should users object to moderation decisions.**[49]

RECOMMENDATIONS TO **STATES**

> Require platforms to provide granular and standardized data to researchers and regulators on **the notification of users** for content deleted, delayed, downgraded, labeled or demonetized, stating:
- ♦ whether the user appealed the decision;
- ♦ whether or not the decision was automated;
- ♦ if the initial decision of the platform was revised: rates of reversal on appeal should be made public;[50]
- ♦ the specific provision of the rules that the content violated, and why the content was thought to violate that provision (including a link to that provision and to the enforcement guidelines related to that specific provision).[51]

Content decisions involving state actors deserve detailed justification precisely because of their impact on human rights and the public interest.

1.3.c TRUSTED FLAGGERS

As noted by the Internet Society, 'over the past few years, we have observed the reliance of some online platforms on 'trusted flagger' initiatives. Even though there is a process of prescribing an entity as a 'flagger', it is still unclear who these 'flaggers' are and how they operate. This lack of clarity inevitably raises questions about transparency and potential conflicts of interest; but, more importantly, it is problematic to outsource quasi-judicial assessments to trusted flaggers, who potentially lack the high standards of due process'.[52]

49 Annenberg Public Policy Center of the University of Pennsylvania (2020). Freedom and Accountability A Transatlantic Framework for Moderating Speech Online. Retrieved from https://www.annenbergpublicpolicycenter.org/feature/transatlantic-working-group-freedom-and-accountability/ (Accessed on 3 August 2020).
50 Bradford, B., et al. (2019). *op. cit.*
51 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
52 Internet Society (2020). *DSA Open Consultation Response*. Retrieved from https://isoc.app.box.com/s/jqjxjqva197iui6u95cciiaz25mgi0x6 (Accessed on 18 October 2020).

RECOMMENDATIONS
TO **STATES**

> **Require platforms to share information describing the procedures for cooperating with 'trusted flaggers': a list of flaggers, selection procedures, privileges attached to the status, statistical data on the number of reports examined, the number of items detected proactively, the follow-up (removal, maintenance, etc.), the appeals processed, and so on.**[53]

1.4. ALGORITHMS & CONTENT MODERATION, RANKING, TARGETING

Understanding the algorithmic mechanisms underlying content moderation, content ranking, content targeting, and social influence building is crucial to assess the dissemination and amplification of problematic content, the influence on the public debate and the formation of the public opinion.

RECOMMENDATIONS
TO **STATES**

> **Require platforms to maintain up-to-date reference documents and release them to vetted researchers and regulators, on each core function of the algorithms,** including ranking (how they rank, organize and present user-generated content), targeting (how they target users with unsolicited content, usually as a paid service, at their own initiative or on behalf of third parties), moderation[54] and social recommendations, as well as detection of content.[55]

> **Require platforms to explain the objectives of the algorithms' optimization.**

> **Require platforms to provide clear information, including granular and standardized data when possible, to explain the:**

- ♦ number of times content has been curated, moderated and ordered—broken down by action taken;
- ♦ process and algorithm employed to identify content or accounts that violate the rules for user-generated content, advertising content, and advertisement targeting;[56]
- ♦ criteria and data used to train the algorithms to identify, moderate, prioritize and personalize content;
- ♦ collection of data, including personal data;
- ♦ use of data in the algorithms' input parameters;
- ♦ biases of the data;
- ♦ processing flow followed;
- ♦ algorithmic components used;
- ♦ supervisory/monitoring mechanisms used in algorithmic learning;

53 *Creating a French Framework. op. cit.*
54 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
55 Contribution received from Institut Montaigne.
56 Maréchal, N. & E. Roberts Biddle. (2020). *op. cit.*

- ♦ personalization carried out;
- ♦ biases reproduced by the algorithms;[57]
- ♦ potential misuses and abuses of the algorithms;
- ♦ false positive / false negative rates;
- ♦ metrics to report on the algorithms' performance;
- ♦ confidence interval applied to the result;
- ♦ procedures used to correct errors;[58]
- ♦ underlying data upon which published estimates of errors are based.[59]

> **Require platforms to describe how many moderators they have, describing in detail their professional profile** (experience, specialization or knowledge), their spatial location and their distribution of tasks (in terms of themes, geographical areas, etc.) without detriment to the respect for the right to privacy and anonymity of moderators.[60] (see chapter 2)

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Make clear to users that content suggestions are the result of editorial choices, which can change.[61]**

> **Publish detailed information on the mechanisms employed during the COVID health crisis.**

> **Enable users to decide whether to allow these algorithms to shape their online experience, and to change the variables that influence them. See Chapter 3 on Platform Design.[62]**

1.5. DISCLOSURE OF REACH OF CONTENT: CONTENT RANKED BY REACH

According to Avaaz's contribution to the working group, at the peak of the first wave of the pandemic in April 2020 content from the web sites of the top ten health misinformers reached an estimated 420 million views—more than four times the number of estimated views garnered by equivalent content

57 In addition, algorithms, especially machine-learning algorithms, may exacerbate unintended biases that are not known by the company itself and thus are not captured and disclosed under an outward transparency scheme. These biases can often only be revealed by an active scrutiny review.
MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*

58 *Creating a French Framework*. *op. cit.*

59 This applies to the algorithms that are used for initial screening, as well as the algorithms that are used to identify content that is likely to violate platform rules. It also applies to initial human reviews, further reviews as requested by complaining users or users whose content has been deleted or downgraded, and the reviews of samples of moderated content that are used to establish an internal baseline of accuracy. The platforms should develop a mechanism to make disaggregated data on the prevalence of violating content available to third-party researchers.
MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*

60 Observacom. (July 2020). A Latin American Perspective. *op. cit.*

61 Comité National Pilote d'Éthique du Numérique. (2020). Enjeux d'éthique dans la lutte contre la désinformation et la mésinformation. Retrieved from https://www.ccne-ethique.fr/sites/default/files/cnpen-desinformation-2020-07-21-2.pdf (Accessed on October 28, 2020).

62 Maréchal, N. & E. Roberts Biddle., *op. cit.*

from the top ten official health information authorities, such as the WHO and the CDC.[63]

At present, nobody outside the social networks companies knows with any certainty such information as what posts/content reaches the highest number of users.[64] This is critical information, especially, for example, ahead of an election, or in the middle of civil unrest. The Facebook tool CrowdTangle is the only tool that currently offers a first level of transparency. It should be improved and replicated by all dominant platforms.

Audiences figures for television channels, radio shows, newspapers and even cinema are regularly made public. It is time to allow citizens access to such information about digital news sources today.

RECOMMENDATIONS
TO **STATES**

> Require all dominant platforms **to disclose daily the top-performing content reaching the highest number of users (including public posts, videos, in which groups, which channels, that day in the country, per language and dialect).**

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Add reach of content in the Facebook tool CrowdTangle.

> **Involve civil society, researchers, journalists and regulators in the development of tools like CrowdTangle for all dominant platforms.**

1.6. DISCLOSURE OF ADVERTISING: A PUBLIC ADVERTISEMENT DATABASE

Expanding transparency requirements for all advertisements would overcome the problem of defining 'political advertisements' and 'political issues', and would reflect that disinformation is broader than the political spectrum. Definitions of political campaigning differ in each country, and finding a shared definition would be an exceedingly difficult process that would fail to take into account national differences and sensitivities.[65]

Such disclosure would create **a universal public advertisement database,**[66] and enable regulators and researchers to audit it. This database needs to be easily searchable by topic, keywords, languages and countries and could be accessible through an application programming interface (API).

63 The research focused on actors spanning across mainly five countries—the United States, the United Kingdom, France, Germany, and Italy.
64 Marantz, A., *op. cit.*
65 European Partnership for Democracy. (March 2020). Virtual Insanity? The Need to Guarantee Transparency in Digital Political Advertising. Retrieved from https://epd.eu/wp-content/uploads/2020/04/Virtual-Insanity-synthesis-of-findings-on-digital-political-advertising-EPD-03-2020.pdf (Accessed on 6 September 2020).
66 Our recommendation is an expansion to one made in the report :
Posetti, J. & K. Bontcheva. (2020). Freedom of Expression and Addressing Disinformation on the Internet.
Chapter 8 : Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Broadband Commission.
Retrieved from https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf (Accessed on 19 October 2020). p. 262.

RECOMMENDATIONS
TO **STATES**

> Require all dominant platforms to **disclose in real time the advertisements**[67] viewed on their services,[68] with information on the:
- ♦ content of the advertisement: a copy of the advertisement, and names of personalities and issues involved;
- ♦ advertiser, including its contact information, location and source of payment;
- ♦ size of the target audience, as well as the number of views the advertisement receives, together with user engagement beyond viewing the advertisement;
- ♦ selection criteria for targeting recipients (while protecting privacy) as is communicated to advertisers, including the data source, inferred profile, lookalike audiences, custom audiences, and A/B testing practices.[69] Platforms should provide clear information, including granular and standardized data, on how they market advertisements to buyers and how they target them to users;
- ♦ date and time of publication, display, and duration;
- ♦ rates charged;[70]
- ♦ revenues from targeted advertising.

> Require all dominant platforms to make advertisements available within 24 hours of publication, maintain access going back ten years, and create programming interfaces to allow long-term studies.[71]

> **Require all dominant platforms to make key disclosures together with the advertisement itself on the platform**, including a visually prominent banner that the ad has been placed by a certain political actor, the party responsible for funding the ad, etc.[72]

1.7. INFORMATION ON THE USE OF USERS' DATA

'User information' is any data that is connected to an identifiable person, or may be connected to such a person by combining datasets or making use of algorithmic data-processing techniques.[73]

'Our data are an extension of our individuality… Transparency is key. For too long, Internet firms have gotten away with hiding from users how much data they collect, what tools and technologies they use to collect data, and the reasons behind their collection of data. We need governments to mandate that Internet firms—and any others that intentionally obscure the real reasons for their data collection from the typically unassuming and uninformed consumer—must detail their data-collection and data-processing practices in clear terms', considers Dipayan Ghosh, director of the Digital Platforms and Democracy initiative at the Harvard Kennedy School.[74]

67 Recommended by AlgorithmWatch in their contribution to this Working Group.
68 For instance, in the United States, the US Congress should consider expanding the scope of the Honest Ads Act to include all advertisements.
69 European Partnership for Democracy. *op. cit.*
70 This requirement should answer the lack of transparency in transaction details and prices highlighted in the Interim Report on the Evaluation of Competition in the Digital Advertising Market, shared with this working group by Jun Murai.
71 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
72 Ghosh, D. (2020). *Terms of Disservice.* Washington, DC: Brookings. p.227.
73 Maréchal, N., R. MacKinnon, & J. Dheere. (2020). Getting to the Source of Infodemics: It's the Business Model. Key Recommendations for Policymakers. *New America*. Retrieved from https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/key-recommendations-for-policymakers (Accessed on 1 September 2020).
74 Ghosh, D. *op. cit.* p. 204-205.

RECOMMENDATIONS
TO **STATES**

> Require platforms to provide clear information to users in relation to their policies on collecting, storing, retaining, using, and sharing users' data, including:

- what data is collected and for what purpose;
- how data is stored and for how long;
- what use is made of users' data;
- with whom users' data is shared, under what conditions, and whether the platforms were paid to do so.

> Require platforms to allow users to obtain all user information pertaining to them (collected and inferred) that the company holds, in a structured data format to allow practical data portability and interoperability.

> Delete all user information within a reasonable time frame after users terminate their account, or at the user's request.[75]

An increasing number of experts support 'data portability' and 'interoperability'. This would require companies to allow their users to pack up their data and take it to a different platform (similar to keeping your cell phone number when you switch operators), thus opening up their platforms to competitors (i.e., 'the right to relocate'), as explained by Ronald Deibert in his recent book *Reset*.[76]

1.8. MANDATORY HUMAN RIGHTS IMPACT ASSESSMENT

'Move fast and break things' is the much-used Silicon Valley saying. 'Prioritizing scale over efficiency in an environment of uncertainty,'[77] is how the co-founder of Linked-In, Reid Hoffman, and the author Chris Yeh define even more precisely the phenomenon of *blitzscaling*: the growth approach of Silicon Valley startups.

This approach should be reconsidered when the mistakes involved can harm our democracies. Requiring human rights assessments from online service providers is one tool for attempting to limit the risk of harm.

Requiring human rights assessments is the only ex-ante transparency requirement. All the other transparency requirements relate to what already happened on the platforms.

Requiring ex-ante human rights assessments introduces a precautionary approach, similar to the precautionary principle of the food or pharmaceutical industries, as discussed in Chapter 3.

As pointed out in Chapter 2, the United Nations Guiding Principles on Business and Human Rights articulate the corporate responsibility to respect human rights. This includes conducting ongoing human rights due diligence in areas where they operate, so as to be able to avoid adverse human rights impacts.[78]

75 Maréchal, N., R. MacKinnon, & J. Dheere, *op. cit.*
76 Deibert, R. (2020). *Reset: Reclaiming the Internet for Civil Society*. Toronto: House of Anansi Press. p. 312.
77 Hoffman, R. & Yeh, C. (2020). *Blitzscaling*. Retrieved from https://www.linkedin.com/learning/reid-hoffman-and-chris-yeh-on-blitzscaling/why-blitzscale (Accessed on 15 October 2020).
78 *United Nations Guiding Principles on Business and Human Rights* (UNGPs). (2011). United Nations. Principle 17(c),

Platforms cannot cite the problem of scale to avoid this responsibility.[79] A human rights due diligence process should identify, prevent, mitigate and account for how platforms will address their impact on human rights.[80]

RECOMMENDATIONS TO **STATES**

> Require platforms to disclose information for the country where they operate about :

- **the adverse human rights impact not only on the users of the platform, but also on rights-holders, of the platform's terms of service, community guidelines, content moderation practices, redress mechanisms, use of users' data, targeted advertising, and algorithmic systems at each significant change, (see Chapter 2). Reporting should be formal, systematic, and comparable;[81]**
- the **social impact** of their algorithms, which could follow the principles for Accountable Algorithms and be presented at the design stage, pre-launch and post-launch;[82]
- the **environmental, social, and governance** (ESG) impacts of their products and policies;
- the **mental health impact** of their products and policies;
- the **civil rights impact**[83] of their products and policies, taking all communities into consideration, as well as the potential for radicalization and hate;[84]
- the **political impact** assessing whether a platform's products, policies or practices regarding political advertising arbitrarily limit the ability of candidates or parties to disseminate their messages (see Chapter 2).[85]

> Require platforms to submit these detailed assessments to regulators and vetted researchers to allow independent audits, and publish summary results on a publicly accessible website.[86]

79 UNGPs, *op. cit.* Principle 14. ('The responsibility of business enterprises to respect human rights applies to all enterprises regardless of their size, sector, operational context, ownership and structure.')
80 See UNGPs. *op. cit.* Principle 15(b).
81 Maréchal, N., R. MacKinnon, & J. Dheere. *op. cit.*
82 Diakopoulos, N., et al. (n.d.). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. Retrieved from https://www.fatml.org/resources/principles-for-accountable-algorithms (Accessed on 19 October 2020).
83 Facebook published its first civil rights audit in July 2020: Facebook's Civil Rights Audit—Final Report. Retrieved from https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf (Accessed on 19 October 2020).
84 Stop Hate for Profit. (2020). Recommended Next Steps. Retrieved from https://www.stophateforprofit.org/productrecommendations (Accessed on 15 October 2020).
85 Organization for Security and Cooperation in Europe (OSCE). (30 April 2020). *Joint Declaration on Freedom of Expression and Elections in the Digital Age*. Para. 2(a)(iii).
86 In the United States, the current Securities Exchange Commission Privacy Impact Assessment could serve as a base model for the newly required Human Rights Assessments: See, https://www.sec.gov/oit/privacy-impact-assessments.

2. THE GOVERNANCE OF TRANSPARENCY

An accountability regime should hold platforms to their promises. Only a public regulator could supervise the implementation of the transparency framework, audit compliance, and take sanction against repeated failures.[87]

2.1. GENERAL PRINCIPLES

The goal of this new approach to legally binding transparency of digital platforms is to provide democratic oversight. **The approach should include democratic safeguards against potential malpractices of the governments and/or regulators themselves.** As a precondition of legally binding transparency, **public regulators should be independent and adhere to human rights principles.** Democratic safeguards are critical to protect regulators from the interests of the political leadership for example.

RECOMMENDATIONS
TO **STATES**

> Governments should make available publicly and in a regular manner:
- **the number and nature of content restrictions, as well as the categories of personal data that they requested from the online service providers, and the reasons and legal frameworks that justified such requests;**
- a clearly defined legal basis for their request;
- responses from the companies and exact steps that were taken as a result of their requests;
- agreements made with online service providers.

> Governments should:
- **strengthen the governance of the national regulator;**
- **reinforce the regulator's independence from the executive branch and accountability to the legislative branch of government;**
- **clarify the appointment process of the national regulator with a clear vetting process.**

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Include in their transparency reporting to regulators all content-related requests issued to them.[88]

87 Annenberg Public Policy Center of the University of Pennsylvania. (2020). *Freedom and Accountability A Transatlantic Framework for Moderating Speech Online.* Retrieved from: https://www.annenbergpublicpolicycenter.org/feature/transatlantic-working-group-freedom-and-accountability/ (Accessed on 3 August 2020).

88 Pírková, E. & Pallero, J. (2020). 26 Recommendations on Content Governance. *op. cit.*

2.2. AUDIT OF TRANSPARENCY REQUIREMENTS

The information provided by online service providers must be open to audit by the appropriate regulator and/or by an independent auditor.[89] Researchers and regulators should have access to platform data in order to audit the systems involved in order to assure the public that the platforms are operating as intended and without unintended bias[90].

Researchers and regulators do not operate at the same level. An audit by regulators should allow democratic oversight of online service providers' practices. Regulators may be assisted by vetted expert researchers[91] in this endeavor, with the researchers acting as deputies of the regulation authorities and/or following other complementary academic purposes. In addition, vetted researchers could conduct independent academic research. These should be rigorously constructed in order to avoid any misuse of the metadata and to guarantee differential privacy methods, in order to protect individual users' data. The Tony Blair Institute for Global Change makes precise recommendations on how a new independent tier of regulatory audit could help solve the information asymetry problem.[92]

RECOMMENDATIONS TO **STATES**

> Require major online service providers to:

- **develop, at their own cost, a secured platform enabling accredited outside researchers to access the data necessary to implement research of general interest;**
- implement the requisite data-processing; and
- extract the results without compromising users' private data or the value of the aggregate data of the social network.

> Mandate an independent regulator to:

- **define priorities for research of general interest, following public consultation and based on the policy dialogue on substantive issues arising from social network operations;**
- organize the process through which academics can apply for access to the platform. The platform itself should not decide on the merits of the research considered, but this decision should be made by peer review committees following academic standards and set up under the supervision of the regulator. The platforms should have the opportunity to comment on the proposed research project;
- settle disputes between platforms and academics that arise from the implementation of this controlled access.[93]

At this time, we do not propose processes or criteria for deciding which researchers or research organisations are qualified. Nevertheless, how this vetting process is done, by whom, and with what criteria is of central importance and must be explored in further detail.

89 In line with contributions received from Institut Montaigne, Algorithm Watch and Institute for Strategic Dialogue.
90 Institute for Strategic Dialogue Joint paper coordinated by Digital action (2020). Algorithm Inspection and Regulatory Access. Retrieved from: https://www.isdglobal.org/wp-content/uploads/2020/04/Algo-inspection-briefing.pdf (Accessed on November 2, 2020)
91 Could include vetted civil society representatives.
92 Beverton-Palmer, M. & R. Beacon. (2020). Online Harms: Bring in the Auditors. Tony Blair Institute for Global Change. Retrieved from https://institute.global/policy/online-harms-bring-auditors (Accessed on 19 October 2020); and Beverton-Palmer, M. & R. Beacon. (2020). Analysis: Applying the Principles of Audit to Online Harms Regulation. Tony Blair Institute for Global Change. Retrieved from https://institute.global/policy/analysis-applying-principles-audit-online-harms-regulation (Accessed on 19 October 2020).
93 MacCarthy, M. (2020). *Transparency Requirements. op. cit.*

Many lessons should be learned from Social Science One,[94] a Harvard scientific program, and its partnership with Facebook. Seniors researchers waited years before being given access to the metadata on the effects of social media on democracy and elections.

Another model to adapt would be that of CASD.eu, a public interest group bringing together the State represented by INSEE, GENES, CNRS, École Polytechnique and HEC Paris created by inter-ministerial decree of 29 December 2018[95]. Its main purpose is to organize and implement secure access services for confidential data for non-profit research, study, evaluation or innovation, activities described as 'research services', mainly public.

2.3. THREE-TIER DISCLOSURE

For online service providers, a three-tier system of disclosure would be the most appropriate structure. While respecting privacy and intellectual property concerns, this structure would offer three tiers of information access, providing:

(a) users with platform rules and complaint procedures;
(b) vetted researchers[96] and regulators with access to databases on moderation activity, algorithm decision outcomes, and other information; and
(c) under limited circumstances, such as an investigation, access to the most restricted classes of commercially sensitive data to regulators, and personally sensitive data to researchers approved by regulators.[97]

OVERVIEW OF THREE-TIER DISCLOSURE

	Public users	**Vetted researchers[98] & regulators**	**Limited access to regulators & researchers approved by regulators**
Nature of information			
Terms of service/ Community guide-lines/internal policies/ enforcement guide-lines/notification of changes	Yes	Implementation standards for Facebook or equivalents for other digital platforms	
Notice of violations of terms of service (TOS) or laws	User who files a complaint and user whose content was affected	Yes	
Content takedowns, flagged content, accounts disabled & content remaining		Yes	

94 See, https://socialscience.one/.
95 Journal Officiel de la République Française. (2018). Arrêté du 20 décembre 2018 portant approbation de la convention constitutive du groupement d'intérêt public « Centre d'accès sécurisé aux données ». Retrieved from https://www.casd.eu/wp/wp-content/uploads/joe_20181229_0301_0053.pdf (Accessed on October 28, 2020).
96 Civil society representatives are included in the generic term 'vetted researchers'.
97 Annenberg Public Policy Center. *op. cit.*
98 Civil society representatives are included in the generic term 'vetted researchers'.

Notification of users & redress mechanism	If their content is removed or downgraded, and to explain redress procedures	Yes	
Trusted flaggers	Yes	Yes	
Algorithms & content moderation	Data points used to make recommendations	Objectives of algorithms	Yes
Disclosure of content reach: a public database of content ranked by reach	Yes	Yes	
Disclosure of advertising: a public advertisement database	Yes	Yes	
Information on use of users' data	Yes	Yes	
Mandatory human rights assessments	Summary	Yes	

2.4. REGIONAL & NATIONAL TRANSPARENCY REGULATION MODELS

Whatever the region or country, this transparency regulation model should be carefully crafted to prevent abusive regulator coercion of platforms, or coercion by political actors working for the political ends of the regulator itself, or by the political party that happens to be in government. The regulator should be prohibited from reaching into the activities of the platforms to direct or dictate a political outcome or to gather intelligence to be used to favor some political actors over others.[99]

Any new regulation model should include a regular impact assessment, as recommended by the Internet Society.[100]

Parliamentary diplomacy should be encouraged in order to coordinate regulation models between states.[101]

99 MacCarthy, M. (2020). *Transparency Requirements*. *op. cit.*
100 Internet Society. *op. cit.*
101 Contribution received from Kamel Ajji, Affiliate Fellow, Yale Law School Information Society Project.

2.4.a EUROPEAN TRANSPARENCY REGULATION MODEL

At the European level, transparency requirements for platforms could be established through a unique European direct regulation (not as a directive), after public and open consultation, proposed by the EU Commission and adopted by the EU Parliament & Council, the European Union co-legislators.

National regulators, acting together as an EU network, should implement it.[102] Individually, each regulator should assess platforms' compliance in their respective countries.

The EU governance of the network of regulators should include the necessary checks and balances to ensure consistency of the regulation across member states and prevent potential abuse of regulation by national regulators. Specifically, **potential sanctions for platforms' non-compliance should be decided or subject to a review at EU level.**

For example, if the French regulator CSA considered that a certain platform hadn't complied with the European transparency requirements, the CSA could consider that this platform be sanctioned. The sanction requested by the French regulator CSA would be validated at the European level, or subject to a check-and-balance mechanism at this level, to ensure consistency of regulatory decisions across the EU and mitigate any political risk.

This could be done by the EU Commission itself or a collegial organization of regulators, such as an updated version of the existing European Regulators Group for Audiovisual Media Services (ERGA[103]).[104] The ERGA will probably be better positioned to play this role and would also maintain the independence of the regulators.

To ensure an ethical system of checks and balances of the national regulators, national regulatory bodies should not take a decision or make a recommendation:

- without national and European public consultation involving the civil society; and/or
- without a referral for opinion to the national privacy regulator; and/or
- without informing their European regulatory peers and/or the EU Commission and/or an independent president of the European regulators.

The adoption of this model would require reforming existing national regulatory systems by:

- expanding the competency of audiovisual media regulators to online service providers;
- empowering national regulators to access a new level of transparency with the online service providers;
- creating strong mechanisms to allow an ongoing engagement with civil society (academics and NGOs), especially in advising national regulators and assisting in the audit of information provided by the digital platforms.

In addition, a close cooperation with the network of national electoral regulators could be useful regarding political parties' transparency requirements, campaigning regulations, and issues of political advertising.[105]

If such a model of transparency was first adopted in Europe, it could spread organically and have an immediate de facto impact in the rest of the world.

---

102 That is, to participate in the policy dialogue between the platforms and civil society, and organize the controlled access of academics to platforms' data to assess the impact of social networks' algorithms on our societies.

103 See, https://erga-online.eu/

104 Discussion with Benoît Loutrel, rapporteur of the report: *Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision.*

105 European Partnership for Democracy. *op. cit.*

2.4.b US TRANSPARENCY REGULATION MODEL

In the United States, a number of regulation models towards more transparency of digital platforms are currently being discussed.

The Platform Accountability and Consumer Transparency (PACT) bill, introduced by Senators Schatz (D-HW) and Thune (R-SD)[106] in June 2020, aims to update Section 230 of the Communications Decency Act of 1996 that provides immunity to online platforms from civil liability based on third-party content and for the removal of content in certain circumstances. The new bill **includes transparency requirements for platforms on their content moderation practices and a complaint mechanism, and it mandates the US National Institute of Standards and Technology (NIST) to develop a voluntary framework for moderation practices, as well as a process for researchers to access data from online service providers.**

These specifics align in many ways with our recommendations and could be an interesting starting point. This would create **a transparency regime supervised by the Federal Trade Commission (FTC),** while not allowing the FTC 'to review any action or decision by a provider of an interactive computer service related to the application of the acceptable use policy of the provider'.[107]

In addition to this transparency regime supervised by the FTC, the Georgetown University Professor Mark MacCarthy recommends the **creation of: 'a dispute resolution system administered by an industry self-regulatory organization aiming to ensure consistency with a company's publicly disclosed content standards...** The Financial Industry Regulatory Authority (FINRA), the self-regulatory organization established to oversee broker-dealers, provides one model.... FINRA is a not-for-profit self-regulatory organization authorized by federal law to regulate the broker-dealer industry. The Securities and Exchange Commission (SEC) supervises FINRA's operations and a board of governors, consisting of an equal number of public and industry representatives, governs it.'[108]

In addition to this transparency regime and dispute resolution system, **a new 'Digital Platform Agency'** could be created, as recommended by Tom Wheeler, the former chairman of the Federal Communications Commission (FCC), to protect consumers and competition. This new Digital Platform Agency 'should be governed by a new congressionally established digital policy built around three concepts:

1. Oversight of digital platform market activity on the basis of risk management rather than micromanagement; this means targeted remedies focused on market outcomes and thereby avoids rigid utility-style regulation,
2. Restoration of common law principles of a duty of care and a duty to deal as the underpinning of DPA authority, and
3. Delivery of these results via an agency that works with the platform companies to develop enforceable behavioral codes while retaining the authority to act independently should that become necessary.'[109]

106 Brian Schatz United States Senator for Hawai'i [official website]. (2020). Schatz, Thune Introduce New Legislation To Update Section 230, Strengthen Rules, Transparency On Online Content Moderation, Hold Internet Companies Accountable For Moderation Practices. Retrieved from https://www.schatz.senate.gov/press-releases/schatz-thune-introduce-new-legislation-to-update-section-230-strengthen-rules-transparency-on-online-content-moderation-hold-internet-companies-accountable-for-moderation-practices (Accessed on 19 October 2020).
The question of platform liability was purposely not addressed in this report.

107 MacCarthy, M. (2020). A Dispute Resolution Program. *op. cit.*

108 Ibid.

109 Wheeler, T., P. Verveer, & G. Kimmelman. (2020). New Digital Realities; New Oversight Solutions. Retrieved from https://shorensteincenter.org/new-digital-realities-tom-wheeler-phil-verveer-gene-kimmelman/ (Accessed on September 1, 2020).

SUGGESTED US REGULATION MODEL

Focus	Suggested models
Transparency	PACT bill: transparency regime overseen by Federal Trade Commission
Dispute resolutions	FINRA model: create a non-governmental regulator overseen by a federal agency
Consumers & data protection	Create Digital Platform Agency
Competition	Create Digital Platform Agency

2.4.c OTHER NATIONAL AND REGIONAL REGULATIONS

There is an absolute and urgent need to access the same level of transparency and accountability all around the world in order to solve many online content moderation and disinformation issues.

In the absence of a regional democratic structure like the European Union, the challenges will be to harmonize transparency requirements and to prevent politically motivated abuse of national regulations.

In the Americas region, the Organization of American States could help develop such a regional and legal governance. In Africa, the role of the African Union could be worth exploring, as well as the role of the Asia-Pacific Economic Cooperation (APEC)[110] or the Association of Southeast Asian Nations (ASEAN) for Asia, or alternatively regional development banks (DB) could help set up such governance structures on a continental basis (AfricanDB, InterAmericanDB, AsianDB, IslamicDB).

Organizing the actions of national regulators as part of a regional network of regulators seems key. Such networks reinforce the capacity to face global digital platforms and allow the design of meaningful checks and balances to prevent the capture of regulators by local political groups or global actors.

2.5. THE SANCTIONS FOR NON-COMPLIANCE

The sanctions for non-compliance could be financial, in the form of a fine for digital platforms of up to three-four percent of the network's global turnover, proportionate to the severity of the non-compliance.

It is important to note that when the US Federal Trade Commission fined Facebook the unprecedented amount of US$ 5 billion in July 2019 for privacy violations, this didn't make much of a dent on the company. Fines need to be potentially extremely significant, especially in case of recurring non-compliance, in order to be efficient.

In addition, **mandatory publicity for non-compliance could be imposed, in the form of a banner ahead of all advertisements visible by all users on the digital platforms.**

The liability of the CEO of the digital platform should be directly linked to the compliance of the company with the legally binding transparency requirements. The CEO could be required to personally sign off on the transparency requirements reports, attesting that the information disclosed is accurate and complete.

As a last resort—for example, if a platform did not pay the fine—**administrative sanctions could be possible, such as closing the platform's access to the affected country after a court decision.** Such extreme sanctions may be needed for rogue or politically motivated platforms operating on an offshore basis from jurisdictions where international judicial cooperation is extremely slow, if not impossible. These sanctioning powers may be exercised only after formal notice and a court decision.[111]

110 See, https://www.apec.org/
111 *Creating a French Framework*. op. cit.

Chapter 2:
Meta-Regulation of Content Moderation

To safeguard the democratic value of online spaces to all, content moderation must be governed by a set of baseline principles (meta-regulation) that protect democratic values and uphold the human rights and dignity of all persons without discrimination. A human rights approach to content moderation would avoid haphazard decision-making by digital platforms and guard against arbitrary requests from states to remove content.

Contents

INTRODUCTION

At present, **platforms decide the terms of online speech based on rules they themselves make, enforced by human moderators and algorithms that the platforms employ, all with very limited public oversight and accountability**. The *UN Guiding Principles on Business and Human Rights* (UNGPs) impose on business enterprises the responsibility to respect human rights in places where they operate. This means that business enterprises should avoid infringing the human rights of others and should address any adverse human rights impacts with which they are involved themselves.[112]

The UNGPs equally apply to platforms and the business of content moderation.[113] This chapter develops **Human Rights Principles for Content Moderation** that responsible platforms should follow to avoid infringing on the human rights of others when moderating content. At the same time, it reiterates the **state's duty to protect human rights online**. The recommendations in this chapter use **international human rights law as a frame of reference**, while acknowledging the role of regional human rights conventions and the case law of regional human rights bodies and tribunals, as may be applicable, in interpreting international human rights law. National laws, to the extent compatible with international human rights law, are also relevant.[114]

The Human Rights Principles for Content Moderation presented in this chapter can form the **normative content of various forms of (co-)regulation.** On one hand, present self-regulation initiatives should embody these principles if they are genuinely committed to respecting human rights. Many large platforms implement some of the recommendations proposed here. On the other hand, **these principles are ultimately meant for a global (co-)regulatory framework** that is opposed to the current model where platforms can arbitrarily impose global content policies that do not necessarily adhere to international human rights law, without any public oversight, and not necessarily with stakeholder input at all times. These principles should be integrated in any future (co-)regulation to be enforced by public and independent oversight institutions.

This chapter applies to all platforms, including the dominant social media companies at present. Responsibilities should be differentiated depending on the scale and resources of the platform concerned and its influence in a particular country or region. Larger responsibilities should be allocated to companies with greater size, capabilities, and per-country usage to carry out the recommendations outlined in this chapter with respect to areas where they operate.

Meta-regulation of content moderation[115] will not on its own be sufficient in fighting against infodemics. The recommendations provided here must be read in conjunction with the proposals outlined in the other chapters in this report in order to be effective.

For example, content moderation by platforms must always be open to appeal by users whose content is deleted, as highlighted many times in the report (see section 1.3 in chapter 1).

112 *United Nations Guiding Principles on Business and Human Rights* (UNGPs). (2011). United Nations. Principle 11.
113 Gillespie, T., *op. cit.*, p. 21. (Content moderation is not simply 'ancillary' to the operations of platforms, but 'essential, constitutional, definitional'.)
114 For this purpose, States should repeal or amend national laws that are incompatible with international human rights law.
115 Content moderation must be governed by a set of baseline principles (meta-regulation) that protect democratic values and uphold the human rights and dignity of all persons without discrimination.

1.THE GOVERNANCE OF TRANSPARENCY

The Human Rights Principles for Content Moderation (HRPCM) are drawn from the Universal Declaration of Human Rights, as articulated in the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR) (these three are collectively referred to as the 'International Bill of Human Rights'). **The commitment to international human rights law refers to the full range of human rights, including but not limited to the right to freedom of expression,** as additionally protected by international humanitarian law, international criminal law, and international environmental law, among others.[116]

1.1. LEGALITY PRINCIPLE

Social media platforms with a genuine commitment to respect the full range of human rights **must express a policy commitment to international human rights law**, which includes but is not limited to the right to freedom of expression and information. Such policy commitment must continually permeate all units within the company equally—for example, for both policy and engineering teams. For this purpose, platforms must ensure that a human rights expert occupies such a regular position within the company as to be able to influence policy throughout the entire internal operations of the platform, and with the authority to guide content policy and enforcement in all areas where platforms operate.

Under international human rights law, limitations on expression must be clearly defined and set out in a transparent and detailed manner. As applied to social media platforms, this would entail publishing the content policies that govern online speech, the various factors used in moderating content, and how these policies are applied by way of illustration. Specific types of content merit different responses, and companies must explain how other factors such as the identity of the user or the frequency of violations could merit more severe responses.

Further, many countries where social media platforms operate involve users from multiple ethnicities, languages, and cultures. Even within one country, the primary language of various ethnic communities in a given jurisdiction may not always be the national language of that country. Responsible platforms would publish content policies in the dialects used in various communities to ensure that these documents can be understood by all users.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Publish a policy commitment to respect the full range of human rights, approved by the most senior level of the company, and such policy commitment to be reflected in all internal rules and operational policies and procedures.[117]**

116 See, UNGPs. *op. cit.* Principle 12 (Commentary), UNGPs (discussing the need for business enterprises to refer to additional international law standards, such as international humanitarian law, as may be applicable).
117 See, UNGPs. *op. cit.* Principle 16(a), (d), (e).

- Expressly refer to the International Bill of Human Rights, the International Labour Organization's (ILO) Declaration on Fundamental Principles and Rights at Work,[118] and other relevant standards provided by other branches of international law, in each platform's terms of service and content policy.[119]
- Define content policies clearly, concisely, and accessibly, and set them out in a transparent manner.[120]
- Publish content policies in both the national language and dialects used in the country where platforms operate.
- Conduct periodic reviews of the effect on human rights of content policy and moderation practices, not just on users of the platform, but also on rights-holders.[121]

Infodemics often jointly involve misinformation and hate speech. In this connection, international human rights law provides a universal framework for defining problematic content. For one, it expressly prohibits any propaganda for war, as well as any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.[122] Furthermore, UN bodies and mandates have also developed working definitions for terms such as 'hate speech' and 'terrorist speech' that responsible platforms should adopt. For this purpose, **platforms should closely follow the relevant work of United Nations bodies such as the Human Rights Council, the Office of the High Commissioner for Human Rights, United Nations Educational, Scientific and Cultural Organization (UNESCO), the resolutions of the UN General Assembly, and Human Rights Committee jurisprudence; and seek guidance in specific areas from the work of international institutions such as UNICEF, UNHCR, and the World Health Organization and international tribunals such as the International Court of Justice and International Criminal Court.**[123]

Hate speech defined

Although there is no international legal definition for 'hate speech', the United Nations Strategy and Plan of Action on Hate Speech, which aims to address the problem of hate speech at the national and global level, uses the following working definition:

'any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.'[124]

118 For the labor issues affecting content moderation, see :
Roberts, S. T. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media;
Gillespie, T. *op. cit.*, pp. 111-140; Retrived from: https://yalebooks.yale.edu/book/9780300235883/behind-screen
Barrett, Paul M. (June 2020). *Who Moderates the Social Media Giants? A Call to End Outsourcing*. NYU Stern Center for Business and Human Rights. Retrieved from https://bhr.stern.nyu.edu/tech-content-moderation-june-2020.

119 See, UNGPs. *op. cit.* Principle 12.

120 For a detailed discussion, see Chapter 1 on Transparency.

121 See, United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 42(a) (recommending that companies conduct periodic reviews of the impact of company products on human rights). See also, Mandatory Human Rights Impact Assessments in Chapter 1 on Transparency.

122 *International Covenant on Civil and Political Rights*. (ICCPR), United Nations. Article 20.

123 Domino, J. (2 January 2020). How Myanmar's Incitement Landscape can Inform Platform Regulation in Situations of Mass Atrocity. *Opinio Juris*. Retrieved from http://opiniojuris.org/2020/01/02/how-myanmars-incitement-landscape-can-inform-platform-regulation-in-situations-of-mass-atrocity/ (Accessed on October 12, 2020).

124 See: https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

According to the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression ('UN Special Rapporteur on freedom of expression'), hate speech policies must be crafted by considering the kinds of interference users may face on the platform. This would require 'noting the legitimacy of restrictions to protect the rights of others'.[125]

Terrorism defined

The UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism ('UN Special Rapporteur on human rights and counterterrorism') developed a model definition of terrorism as follows:

Terrorism means an action or attempted action where:

1. The action:
   (a) Constituted the intentional taking of hostages; or
   (b) Is intended to cause death or serious bodily injury to one or more members of the general population or segments of it; or
   (c) Involved lethal or serious physical violence against one or more members of the general population or segments of it;

and

2. The action is done or attempted with the intention of:
   (a) Provoking a state of terror in the general public or a segment of it; or
   (b) Compelling a government or international organization to do or abstain from doing something;

and

3. The action corresponds to:
   (a) The definition of a serious offence in national law, enacted for the purpose of complying with international conventions and protocols relating to terrorism or with resolutions of the Security Council relating to terrorism; or
   (b) All elements of a serious crime defined by national law.

Incitement to terrorism defined

The UN Special Rapporteur on human rights and counterterrorism defines incitement to terrorism as follows:

'It is an offence to intentionally and unlawfully distribute or otherwise make available a message to the public with the intent to incite the commission of a terrorist offence, where such conduct, whether or not expressly advocating terrorist offences, causes a danger that one or more such offences may be committed.'[126]

In addition to terrorist content, platforms increasingly take down accounts linked to individuals and organizations that are considered 'terrorist' or 'dangerous' according to privately drafted terms of service. Platforms must depend on civil society input or legitimate judicial processes in the takedown of these accounts. In particular, removal of terrorist organizations must allow for independent judicial review.[127]

125 Ibid., para. 47(b) ('For example, companies could consider how hateful online expression can incite violence that threatens life, infringes upon the freedom of expression and access to information of others, interferes with privacy or the right to vote and so forth. The companies are not in the position of Governments to assess threats to national security and public order, and hate speech restrictions on those grounds should be based not on company assessment but on legal orders from States, themselves subject to the strict conditions established under article 19 (3) of the [ICCPR]').

126 See para 32: https://undocs.org/A/HRC/16/51

127 Scheinin, M. (2010). *Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism. Ten Areas of Best Practices in Countering Terrorism*. A/HRC/16/51. para. 35 (Practice 9. Core elements of best practice in the listing of terrorist entities.). UN Human Rights Council. Retrieved from https://documents-dds-ny.un.org/doc/UNDOC/GEN/G10/178/98/PDF/G1017898.pdf?OpenElement

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Align definitions of problematic content, such as 'hate speech' and 'incitement to terrorism', with those developed by public international institutions, such as relevant United Nations treaty bodies, special-procedure mandate holders and other experts, the World Health Organization, and international courts and tribunals.**

> Consider the following factors when evaluating content for hate speech: [128]

- Context: Contextual analysis should involve communities most affected by content identified as hate speech.[129] This includes evaluating the linguistic, cultural, political, social, and historical context of the post. For this purpose, platforms must hire human moderators fluent in the vernacular used in the post.[130]Platforms must also involve affected communities in deciding the most effective tools to address hate speech posted on the platform. Use of automation or artificial intelligence tools should involve human-in-the-loop.[131]
- Status and intent of the speaker: Examining the intent of the user for posting potentially violating content would prevent companies from moderating content that is shared for journalistic or educational purposes.
- Content and extent of dissemination: This would require platforms to look at whether privacy settings for the post are restricted or public.
- Likelihood of imminent harm to users and the public: This would require platforms to assess whether particular content could lead to adverse impacts on human rights not only for certain users of the platform, but also for rights-holders who may not be users of the platform. Instead of platforms defining these terms individually and arbitrarily, platforms should track definitions according to the standards of international human rights law developed by international institutions, which have more developed and inclusive mechanisms for input in place, and are accountable to the public.[132]

> Adopt the broadest range of protected groups for a platform's hate speech policy, by taking into account historically disadvantaged groups in areas where the platform operates globally.[133]

> Develop hate speech policies by considering the kinds of interference users may face on the platform.

> **Consider civil society input before banning particular individuals and organizations from the platform on account of being 'terrorist' or 'dangerous' organizations.**

128 Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence. (2020). United Nations Human Rights Office of the High Commissioner. Retrieved from https://www.ohchr.org/en/issues/freedomopinion/articles19-20/pages/index.aspx;

129 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 58 (e).

130 "Now there are some fifteen thousand, most of whom are contract workers in cities around the world (Dublin, Austin, Berlin, Manila). See : Marantz, A. (2020). Why Facebook Can't Fix Itself. *The New Yorker.* Retrieved from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself?mc_cid=a2705e31cc&mc_eid=dd9bd17d22 (Accessed on 20 October 2020).

131 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 58 (d) and (e).

132 See, e.g.,
United Nations. *UN Strategy and Plan of Action on Hate Speech.* (2019). Retrieved from https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

133 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom, *op. cit.* para. 47 (a).

1.2. NECESSITY AND PROPORTIONALITY PRINCIPLE

International human rights law demands that limitations on expression should only be that which is necessary to achieve a legitimate purpose.[134] Restrictions should not be overbroad; the least intrusive means should be used to limit expression.[135] International human rights law also distinguishes between public figures and private individuals, with more protection for the latter's right to freedom of expression and information. With respect to public figures, there must be further distinction between state actors, on one hand, and journalists, human rights defenders, and activists that might attain the status of a public figure in a given country or context.

TYPES OF USERS

Not only individual users, but also state actors have used platforms to disseminate hate speech constituting incitement to discrimination, hostility, or violence under international law, or to wage disinformation campaigns against human rights defenders, members of civil society, and members of minority groups. According to the UN Special Rapporteur on freedom of expression, **politicians, government and military officials, and other public figures, 'should be bound by the same hate speech rules that apply under international standards', due to their 'prominent and potential leadership role in inciting behavior'**.[136]

A 'public figure' standard should guide moderation of content posted by state actors insofar as it should require a different response to content found to be in violation of the definition of hate speech proposed above.[137] For this purpose, platforms must distinguish between public figures and state actors on one hand, and private individuals, on the other hand, when deciding on content responses. Platforms already treat different categories of users differently with respect to certain content. Content posted by state actors has been taken down for violating platforms' COVID-19-related misinformation policies.[138] Politicians have also been exempted from the application of certain content policies.[139]

In addition to state actors, religious leaders, celebrities, and other people who can influence public discourse due to their large audience in a particular socio-political context, should also be included.

TYPES OF CONTENT

Platforms moderate different types of content:

a. **Illegal**: Is the content incompatible under international human rights law or national law pursuant to a legitimate court process?
b. **Legal yet harmful**: Apart from moderating illegal content such as child pornography or incitement to violence, platforms also moderate content that is legal yet deemed harmful by the platform.

---

134 ICCPR. *op. cit.* Article 19(3).
135 United Nations. UN Human Rights Committee, General Comment no. 34: Article 19: Freedoms of Opinion and Expression, CCPR/C/GC/34 (12 September 2011), para. 34.
136 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 47(d).
137 By analogy, under international human rights law, public figures have less protection in defamation suits. See Domino, J. (2020). The Facebook Oversight Board as Operational Level Grievance Mechanism. Cambridge Core Blog. Retrieved from https://www.cambridge.org/core/blog/2020/04/01/the-facebook-oversight-board-as-operational-level-grievance-mechanism/ (Accessed on 12 October 2020).
138 See, e.g.
Brito, C. (2020). Facebook, Twitter and Youtube Take Down False Coronavirus 'Cure' Video Shared by Trump. CBS News. Retrieved from https://www.cbsnews.com/news/facebook-twitter-youtube-removing-false-covid-19-information-video-trump-share/ (Accessed on 12 October 2020).
Lyons, K. (2020). Twitter Removes Tweets by Brazil, Venezuela Presidents for Violating Covid-19 Content Rules. *The Verge*. Retrieved from https://www.theverge.com/2020/3/30/21199845/twitter-tweets-brazil-venezuela-presidents-covid-19-coronavirus-jair-bolsonaro-maduro (Accessed on 12 October 2020).
YouTube, COVID-19 Medical Misinformation Policy. *op. cit.*
139 See, e.g.,
Twitter. (2020). About Public-Interest Exceptions on Twitter. Retrieved from https://help.twitter.com/en/rules-and-policies/public-interest (Accessed on 12 October 2020).
Facebook. (2016). Input from Community and Partners on our Community Standards. Retrieved from https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/

An example of this would be TikTok's policy against adult nudity;[140] Facebook's policy against 'inauthentic behavior' or 'cruel and insensitive' posts;[141] or Snapchat's policy against ' impersonation, deceptive practices, & false information'. [142] In evaluating content that is legal under international human rights law yet considered harmful by the platform, input from affected communities and civil society must be considered in deciding whether or not the content should be moderated.[143]

DISTINGUISHING RESPONSES

Platforms have a wide latitude in responding to particular types of problematic content. **Content responses must make use of tools that promote individual autonomy, security and free expression, and involve de-amplification, demonetization, education, counter-speech, reporting, and training as alternatives, when appropriate, to the banning of accounts and the removal of content.**[144] In the words of the UN Special Rapporteur on freedom of expression:

> [Platforms] can delete content, restrict its virality, label its origin, suspend the relevant user, suspend the organization sponsoring the content, develop ratings to highlight a person's use of prohibited content, temporarily restrict content while a team is conducting a review, preclude users from monetizing their content, create friction in the sharing of content, affix warnings and labels to content, provide individuals with greater capacity to block other users, minimize the amplification of the content, interfere with bots and coordinated online mob behaviour, adopt geolocated restrictions and even promote counter-messaging. Not all of these tools are appropriate in every circumstance, and they may require limitations themselves, but they show the range of options short of deletion that may be available to companies in given situations. In other words, just as States should evaluate whether a limitation on speech is the least restrictive approach, so too should companies carry out this kind of evaluation. And, in carrying out the evaluation, companies should bear the burden of publicly demonstrating necessity and proportionality when so requested by affected users, whether the user is the speaker, the alleged victim, another person who came across the content or a member of the public.[145]

The COVID-19 pandemic has exposed the ability of platforms to moderate problematic content when they choose to, and to employ a variety of measures in doing so.[146] The dominant platforms have moderated content that contradicts authoritative health sources globally and locally, such as the World Health Organization, even if such content is posted by a state actor. Twitter, for instance, currently assigns labels to relevant tweets allowing users to refer to external trusted sources of information. It also puts warning messages to specific posts.[147] Platforms should moderate other types of COVID-19-related content, such as discriminatory rhetoric and hate speech towards certain nationalities and members of ethnic, gender, and religious minorities.[148]

---

140 TikTok. (2020). Community Guidelines. Retrieved from https://www.tiktok.com/community-guidelines?lang=en (Accessed on 12 October 2020).
141 Facebook. (2020). Community Standards. Retrieved from https://www.facebook.com/communitystandards/introduction (Accessed on 12 October 2020).
142 Snapchat. (2020). Community Guidelines. Retrieved from https://www.snap.com/en-US/community-guidelines (Accessed on 12 October 2020).
143 Industry standards and best practices can also be consulted, although this should never replace input from affected communities and civil society. See, e.g., Global Alliance for Responsible Media, Brand Safety Floor & Sustainability Framework (providing definition of harmful content for safe advertising practices, among others), World Federation of Advertisers. (2020). WFA and Platforms Make Major Progress to Address Harmful Content. Retrieved from https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content (Accessed on 12 October 2020).
144 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 58(f).
145 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 51.
146 Douek, E. (2020). COVID-19 and Social Media Content Moderation. *Lawfare*. Retrieved from https://www.lawfareblog.com/covid-19-and-social-media-content-moderation (Accessed on 12 October 2020).
147 Twitter. (2020). Updating our Approach to Misleading Information. Retrieved from https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html (Accessed on 12 October 2020).
148 See, United Nations. (2020). United Nations Guidance Note on Addressing and Countering COVID-19-Related Hate Speech. Retrieved from https://www.un.org/en/genocideprevention/documents/Guidance%20on%20COVID-19%20related%20Hate%20Speech.pdf (Accessed on 12 October 2020).

The availability of a broad range of measures should enable platforms to 'tailor their responses to specific problematic content, according to its severity and other factors'.[149]

These factors include the following:

1. Type of content
   - ♦ Illegal
   - ♦ Legal yet harmful

These two categories should be given differentiated responses,[150] depending further on other factors listed here.

2. Speaker

Is the content posted by a state actor or other public figure who can influence public discourse in the area primarily affected by the content?

3. Severity of content

4. Frequency of content policy violation

How many times has the user posted problematic content?
Does the content repeatedly violate the same content policy, or different content policies?

5. Context

Infodemics have a tendency to arise during moments of crisis. Such contexts or events should therefore merit a specialized set of responses for a limited duration. The UNGPs recommend businesses to conduct ongoing human rights due diligence in areas where they operate, so as to be able to address human rights risks as they evolve.[151] Platforms cannot cite the problem of scale to avoid this responsibility.[152] A human rights due diligence process should identify, prevent, mitigate and account for how platforms address their impact on human rights.[153] However, priority should be given to countries where any of the following conditions are present:

- > the platform has an outsized presence;
- > the country in which the platform operates is experiencing significant political developments that may involve a surge in content, including but not limited to national elections, armed conflict, or a natural disaster;
- > potentially violating content is posted by a public figure (e.g., a state actor);
- > where civil society or international institutions flag a particular situation as warranting urgent attention.

For this purpose, platforms should closely follow the relevant work of United Nations bodies such as the Human Rights Council, Office of the High Commissioner for Human Rights, resolutions of the UN General Assembly, United Nations Educational, Scientific and Cultural Organization (UNESCO), and Human Rights Committee jurisprudence; and in specific areas seek guidance from the work of international institutions such as UNICEF, UNHCR, and the World Health Organization and international tribunals such as the International Court of Justice and International Criminal Court.[154]

A risk ratio can be adopted by platforms such that when a particular country goes beyond a certain risk threshold, it would adopt a more aggressive set of content responses within a time-bar. This can include

149 Ibid.
150 See, e.g., Observacom, written contribution to this working group. Para. 4(D) (recommending that for "'offensive", "inappropriate", "indecent" and similar vague or broad definitions, which could illegitimately affect freedom of expression, big platforms should provide mechanisms and notices for other users—voluntarily and based on their moral, religious, cultural, political or other preferences— to decide whether they want to have access to it. Such content should not be prohibited, removed or reduced in scope by default if it passes the test of legality, necessity and proportionality, since doing so would disproportionately affect users' right to freedom of expression.').
151 UNGPs. *op. cit.* Principle 17(c),
152 UNGPs. *op. cit.* Principle 14. ('The responsibility of business enterprises to respect human rights applies to all enterprises regardless of their size, sector, operational context, ownership and structure.')
153 See, UNGPs. *op. cit.* Principle 15(b). See also, Mandatory Human Rights Assessment in Chapter 1 on Transparency.
154 Domino, J. (2 January 2020). How Myanmar's Incitement Landscape can Inform Platform Regulation in Situations of Mass Atrocity. *Opinio Juris*. Retrieved from http://opiniojuris.org/2020/01/02/how-myanmars-incitement-landscape-can-inform-platform-regulation-in-situations-of-mass-atrocity/ (Accessed on 12 October 2020).

additional human moderators devoted to reviewing content, and/or the introduction of various content responses (e.g. introducing friction, imposing a time-bar before content is posted).[155]

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Treat different categories of users differently when moderating content, in line with the "public figure" standard under international human rights law.**

> **Use various tools in content moderation, depending on the type of problematic content and other factors (see discussion above).**

> **Develop tools that promote individual autonomy, security and free expression, and involve de-amplification, demonetization, education, counter-speech, reporting, and training as alternatives, when appropriate, to the banning of accounts and the removal of content.[156]**

> **Track the work of relevant public international institutions when identifying at-risk situations or countries, potentially using a risk ratio approach, based on the following factors:**
- **the platform has an outsized presence;**
- **the country in which the platform operates is experiencing significant political developments that may involve a surge in content, including but not limited to national elections, armed conflict, or a natural disaster;**
- **potentially violating content is posted by a public figure (e.g., a state actor);**
- **where civil society or international institutions flag a particular situation as warranting urgent attention.**

> **Expand the number of moderators and spend a minimal percentage of platforms' income to improve quality of content review, especially in at-risk countries/situations.[157]**

> **Significantly expand fact-checking to address misinformation and disinformation.[158]**

> **Hire a content overseer.[159] Platforms should appoint a senior official taskedto oversee the development of content policies and their execution as well as supervise fact-checking. This person would coordinate all the relevant teams involved in developing and executing policy. Such a person should have expert knowledge in human rights law and work directly or be supervised by the human rights director of the platform.**

155 The Center for Humane Technology is currently developing a risk ratio approach to content moderation. (Source: Interview of lead rapporteur with Tristan Harris, October 2020).
156 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 58(f).
157 Barrett, *op. cit.,* p. 24.
Tristan Harris also recommends increasing the budget for content moderation. See
Gonzales, G. (2020). Increasing Sharing Friction, Trust, and Safety Spending may be Key Facebook Fixes. Rappler.
Retrieved from https://www.rappler.com/technology/features/tristan-harris-aza-raskin-maria-ressa-undivided-attention-podcast (Accessed on November 1, 2020).
158 Barrett, *op. cit.,* p. 26.
159 Barrett, *op. cit.*, p. 25. (although the recommendation here is more prescriptive than Barrett's version).

> **Expand moderation in at-risk countries based on factors outlined above.**[160]

> **Establish a country team in countries where platforms operate, whenever possible.** Platforms should make a priority of locally embedding human moderators and country teams in the various countries where their products and services are used. In certain countries, there may be legitimate security concerns for moderators and employees working in-country. In such instances, platforms should make a good-faith determination that establishing an in-country team would put their moderators and staff at risk, and therefore base them instead in a neighboring state, where the platform maintains other moderators and staff.

> Dominant platforms should share their knowledge and tools widely as open-source to ensure that smaller platforms and markets can have access to the same content-moderation technology.[161]

1.3. LEGITIMACY PRINCIPLE

International human rights law recognizes the following as legitimate aims that could justify limiting expression:

1. Respect for the rights or reputations of others
2. National security
3. Public order
4. Public health
5. Public morals[162]

As applied to platforms, the moderation of COVID-19-related misinformation shows that platforms perform a crucial role in limiting expression inimical to public health.

As discussed in HRPCM Principle 2, above, platforms moderate content that goes beyond the baseline legitimate aims that human rights law provides. For this latter purpose, platforms must have differentiated responses to harmful yet legal content, depending additionally on other factors discussed under Principle 2.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Track assessments of relevant public international institutions or legitimate state assessments when moderating content that has an impact on the rights or reputations of others, national security, public order, public health, or morals.**

160 Barrett, *op. cit.*, p. 25.
161 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 50.
162 ICCPR. *op. cit*. Article 19(3).
United Nations. UN Human Rights Committee, General Comment no. 34: Article 19: Freedoms of Opinion and Expression, CCPR/C/GC/34 (12 September 2011), para. 32: 'The Committee observed in general comment No. 22, that "the concept of morals derives from many social, philosophical and religious traditions; consequently, limitations... for the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition". Any such limitations must be understood in the light of universality of human rights and the principle of non-discrimination.'
Human Rights Committee. (2011). General Comment No. 34. Article 19: Freedoms of opinion and expression. Retrieved from https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf (Accessed on 12 October 2020).

> **Moderate verifiable misinformation based on findings and assessments of public international institutions such as the World Health Organization and relevant UN bodies.**

> **Employ a wide range of tools when moderating content that limits expression beyond the minimum legitimate aims, in accordance with HRPCM Principle 2.**

1.4. EQUALITY PRINCIPLE

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Make a reasonable effort to adopt measures that make it possible for users to access a diversity of political views and perspectives.**

> **Ensure that automated tools, such as algorithmic ranking, do not intentionally, or unintentionally, unduly hinder access to election-related content and the availability of a diversity of viewpoints to users.**[163]

> **Undertake, as part of ongoing human rights due diligence, an assessment of whether a platform's products, policies or practices regarding political advertising arbitrarily limit the ability of candidates or parties to disseminate their messages.**[164]

> Assume the same kind of obligations that broadcasters have in the different jurisdictions where they operate. An example would be the voluntary fairness doctrine in the United States.[165]

Voluntary Fairness Doctrine[166]

The Voluntary Fairness Doctrine aims to guarantee the quality of election-related information that is distributed widely. As applied to platforms, the doctrine would apply broadcaster principles on providing equal access and equal information to users. It would require platforms to take on responsibility for providing the same information to voters at no cost. For example, they could provide equal amounts of campaign advertising for free to different political parties and distribute it before elections. Each item would be labeled as to where it comes from. Ideally, each item would be fact-checked before appearing. But if not, each piece of information should be labeled with a disclaimer that the information within has not been factually verified, and a list of verified websites provided to encourage audiences to use their critical thinking skills and visit reputable websites to check facts that might be fabricated or presented out of context[167]. Platforms should also find a way to make high-quality information salient.

163 See Organization for Security and Cooperation in Europe (OSCE). (30 April 2020). *Joint Declaration on Freedom of Expression and Elections in the Digital Age*. Para. 2(a)(ii).
164 Ibid., para. 2(a)(iii).
165 See, e.g., Schiffrin, A. (2020). *Beyond Transparency: Regulating Online Political Advertising*. Roosevelt Institute. Retrieved from https://rooseveltinstitute.org/publications/beyond-transparency-regulating-online-political-advertising/ (Accessed on 27 October 2020)
166 Ibid.
167 Berger, Guy. 2019. Personal interview by Anya Schiffrin. September 13, 2019.

Other options include:

Providing Free Advertising to Registered Political Parties

A different approach would be to provide free ads equally to registered parties only, since providing free ad slots to other entities promoting various causes might be too demanding.[168]

Abolishing political advertising altogether and instead disseminating equal amounts of news about each campaign's position on key issues before the election. This could include live-streaming of debates or town halls, or transmitting clearly labeled information from the campaigns in prominent locations.

The public-service duties would simply be giving the tech companies a new function, which is to consciously add relevant information about public affairs to the online conversation rather than simply ranking or suppressing it. This use of counter-speech is not dissimilar to what YouTube does with its Creators for Change program, which finds and fosters influencers who can post high-quality videos in order to counter false and inflammatory videos.[169]

1.5. NON-DISCRIMINATION PRINCIPLE

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Apply content policy to similar categories of users without discrimination, in line with the HRPCM Principle 2.**

> **Ensure that algorithms used to enforce content policies do not reinforce discriminatory biases.[170]**

> **Diversify the workforce throughout the entire company.**

> **Conduct regular unconscious-bias workshops, particularly for teams involved in developing content policy and the algorithms for moderating content.**

168 Berger, Guy. 2019. Personal interview by Anya Schiffrin. September 13, 2019.
169 Jahromi, Neima. 2019. "The Fight for the Future of YouTube." The New Yorker, July 8, 2019. https://www.newyorker.com/tech/annals-of-technology/the-fight-for-the-future-of-youtube
170 Observacom, written comment, Para. 3(A) (Recommending that large platforms 'not use discriminatory criteria' in the 'curation/ prioritization of the visualization' of user-generated content [in news feeds, search results, news access services, etc.]).

2. SUMMARY OF RECOMMENDATIONS TO SERVICE PROVIDERS

In sum, responsible platforms must adopt the following recommendations if they are genuinely committed to respecting human rights in the conduct of their business:

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Reaffirm commitment to the UNGPs.**

> **Provide effective remedy to rights-holders affected by moderation of specific content, including but not limited to users of the platform.**[171]

> **Adopt accessible and efficient operational-level grievance and remedy mechanisms, beyond user-appeal procedures.**

> **Adopt the Human Rights Principles for Content Moderation (HRPCM) summarized below:**

PRINCIPLE	RECOMMENDATION
Legality	**Publish a policy commitment to respect human rights, approved by the most senior level of the company, and such policy commitment to be reflected in all internal rules and operational policies and procedures.**
	Mention adherence to the International Bill of Human Rights and the International Labour Organization's (ILO) Declaration on Fundamental Principles and Rights at Work in the platform's terms of service and content policy.
	Define content policies clearly, concisely, and accessibly, and set them out in a transparent manner (see part 1.2 Chapter 1).
	Publish content policies in both the national language and dialects used in the country where platforms operate.
	Conduct periodic reviews of the effect on human rights of content policy and moderation practices, not just on users of the platform, but also on rights-holders (see part 1.8 Chapter 1).

171 See, UNGPs. op. cit. Principle 27.

	Align definitions of problematic content, such as 'hate speech' and 'incitement to terrorism', with those developed by public international institutions, such as relevant United Nations treaty bodies, special-procedure mandate holders and other experts, the World Health Organization, and international courts and tribunals.
	Consider the following factors when evaluating content for hate speech: **(a) Context;** **(b) Status and intent of the speaker;** **(c) Content and extent of dissemination;** **(d) Likelihood of imminent harm to users and the public.**
	Adopt the broadest range of protected groups for the platform's hate speech policy, by taking into account historically disadvantaged groups in areas where the platform operates globally.
	Develop hate speech policies by considering the kinds of interference users may face on the platform.
	Consider civil society input before banning particular individuals and organizations from the platform on account of being 'terrorist' or 'dangerous' organizations.
Necessity and Proportionality	**Treat different categories of users differently when moderating content, in line with the "public figure" standard under international human rights law.**
	Use various tools in content moderation, depending on the type of problematic content and other factors.
	Develop tools that promote individual autonomy, security and free expression, and involve de-amplification, demonetization, education, counter-speech, reporting, and training as alternatives, when appropriate, to the banning of accounts and the removal of content.
	Track the work of relevant public international institutions when identifying at-risk situations or countries, potentially using a risk ratio approach, based on the following factors: **> the platform has an outsized presence;** **> the country in which the platform operates is experiencing significant political developments that may involve a surge in content, including but not limited to national elections, armed conflict, or a natural disaster;** **> potentially violating content is posted by a public figure (e.g., a state actor);** **> where civil society or international institutions flag a particular situation as warranting urgent attention.**

	Expand the number of moderators and spend a minimal percentage of platforms' income to improve quality of content review, especially in at-risk countries/situations.
	Significantly expand fact-checking to address misinformation and disinformation.
	Hire a content overseer.
	Expand moderation in at-risk countries, based on factors outlined under HRPCM Principle 2.
	Establish a country team in countries where platforms operate, whenever possible.
	Dominant platforms should share their knowledge and tools widely as open-source to ensure that smaller platforms and markets can have access to the same content-moderation technology.
Legitimacy	**Track assessments of relevant public international institutions or legitimate state assessments when moderating content that has an impact on the rights or reputations of others, national security, public order, public health, or morals.**
	Moderate verifiable misinformation based on findings and assessments of public international institutions such as the World Health Organization and relevant UN bodies.
	Employ a wide range of tools when moderating content that limits expression beyond the minimum legitimate aims, in accordance with HRPCM Principle 2.
Equality	**Assume the same kind of obligations that broadcasters have in the different jurisdictions where they operate.**
	Make a reasonable effort to adopt measures that make it possible for users to access a diversity of political views and perspectives (see section on the mandatory level of noise, in Chapter 3).
	Ensure that automated tools, such as algorithmic ranking, do not intentionally or unintentionally, unduly hinder access to election-related content and the availability of a diversity of viewpoints to users.
	Undertake, as part of ongoing human rights due diligence, an assessment of whether the platform's products, policies or practices regarding political advertising arbitrarily limit the ability of candidates or parties to disseminate their messages (see Chapter 1).

Non-discrimination	**Apply content policy to similar categories of users without discrimination, in line with the HRPCM Principle 2.**
	Ensure that algorithms used to enforce content policies do not reinforce discriminatory biases.
	Diversify workforce throughout the entire company.
	Conduct regular unconscious-bias workshops, particularly for teams involved in developing content policy and the algorithms for moderating content.

3. RECOMMENDATIONS TO STATES

Even as platforms decide the terms of online speech, States should not forget their role as primary duty-bearers under international human rights law. They should protect human rights online by refraining from violating the human rights of private individuals and regulating platforms in such a way that protects the human rights of users and affected stakeholders.

RECOMMENDATIONS TO **STATES**

> **Reaffirm commitment to the UNGPs, including the duty to protect human rights offline as well as online.**[172]

> **Commit to the Human Rights Principles for Content Moderation (HRPCM) specified above, including in designing laws or regulations concerning online service providers.**

> **Support the establishment of a global governance framework with the mandate to monitor issues arising from technology, including content moderation, and to coordinate the development of standards and best practices to inform national and/or international regulation.**[173]

172 United Nations. (2011). The UN Guiding Principles on Business and Human Rights: Implementing the 'Protect, Respect, Remedy' Framework was Unanimously Endorsed by the UN Human Rights Council in Resolution 17/4. Retrieved from https://undocs.org/en/A/HRC/RES/17/4 (Accessed on 12 October 2020).

173 See, e.g.
Fay, R. (2019). Digital Platforms Require a Global Governance Framework (on Creating a Digital Stability Board). Center for International Governance Innovation.
Retrieved from https://www.cigionline.org/articles/digital-platforms-require-global-governance-framework (Accessed on 12 October 2020).
Council of Europe Parliamentary Assembly. (2020). Towards an Internet Ombudsman Institution.
Retrieved from https://pace.coe.int/en/files/28728/html (Accessed on 12 October 2020).

> Regulate platforms' implementation of responsibilities under the UNGPs, and establish transparency oversight mechanisms, such as the one proposed in part 2 of Chapter 1.[174]

> Provide guidance to platforms for human rights due diligence processes, such as discussed in part 1.8 Chapter 1.[175]

> Repeal any law that criminalizes or unduly restricts expression, online or offline.[176]

> Repeal any law that criminalizes or unduly restricts expression based on vague and ambiguous ideas, such as 'false news' or 'non-objective information'.[177]

> Refrain from demanding of platforms—through legal or extralegal threats—that they take action to moderate content that international human rights law would bar states from taking directly.[178]

> Refrain from establishing laws or arrangements that would require the 'proactive' monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship.[179]

> Refrain from adopting models of regulation where government agencies, rather than judicial authorities, become the arbiters of lawful expression.[180] This includes election commissions that intend to monitor social media campaigns.

> Design regulation that considers the extraterritorial impact of intended regulation on the human rights of users globally.[181]

> Adopt laws that require platforms to create databases of action that companies take against hate speech, and to otherwise encourage companies to respect human rights standards in their own rules.[182]

174 OSCE. (2019). Joint Declaration: Challenges to Freedom of Expression in the Next Decade. Retrieved from https://www.osce.org/representative-on-freedom-of-media/425282 (Accessed on 12 October 2020).
175 Wingfield, R., I. Tuta, & T. Bansal, (2020). *The Tech Sector and National Action Plans on Business and Human Rights*. The Danish Institute of Human Rights. Retrieved from https://www.humanrights.dk/publications/tech-sector-national-action-plans-business-human-rights (Accessed on 12 October 2020). p. 46.
176 UN Human Rights Council. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. A/HRC/38/35 (6 April 2018), para. 65.
177 OSCE.. (2017). Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda. Retrieved from https://www.osce.org/files/f/documents/6/8/302796.pdf (Accessed on 12 October 2020).
178 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 57.
179 UN Human Rights Council. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*,. A/HRC/38/35 (6 April 2018), para. 67.
180 Ibid. para. 68.
181 Wingfield, R. et al.. *op. cit.*, p. 38 ('While the UNGPs say that states are not generally required to regulate the extraterritorial activities of businesses domiciled in their territory and/or jurisdiction, they also recognize that states are not generally prohibited from doing so, providing that there is a recognized jurisdictional basis. The UNGPs recognize that there may be strong policy reasons for states to be clear about their expectations of businesses abroad. States do not have unlimited power to enact laws which apply to extraterritorial activities and must operate within the constraints of international law and comity.... The regulatory framework which applies to companies in one state, particularly their home state, will often have impacts in others in which the company operates. For example, the EU's [General Data Protection Regulation] ... sets higher standards than most other national data protection frameworks. Rather than having many different data protection policies for different states, some tech companies simply use the GDPR requirements as their global data protection policy....').
182 United Nations. UN General Assembly, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General, A/74/486 (9 October 2019), para. 57(f).

> **Publish detailed transparency reports on all content-related requests issued to platforms and involve genuine public input in all regulatory considerations (as detailed in Chapter 1).[183]**

> **Adopt state-based judicial and non-judicial grievance mechanisms to provide remedy to aggrieved users and affected rights-holders.[184]**

> **Adopt laws on election campaigns and advertising to address state-sponsored disinformation online.[185]**

> **Refrain from making, sponsoring, encouraging, or further disseminating statements which they know or reasonably should know to be false (disinformation) or which demonstrate a reckless disregard for verifiable information (propaganda).[186]**

> **Ensure the dissemination of reliable and trustworthy information, including that about matters of public interest, such as the economy, public health, security and the environment.[187]**

> **Support positive measures to address online disinformation, such as the promotion of independent fact-checking mechanisms and public education campaigns.[188]**

183 UN Human Rights Council. (2018) *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.* A/HRC/38/35 (6 April 2018), para. 69.

184 UNGPs. *op. cit.* Principles 25-27. These grievance mechanisms include strengthening existing institutions, including but not limited to the communications and information ministries, national human rights institutions, data protection authorities, and competition commissions, to respond effectively to digital rights issues.

185 See
Ong, J. C. & R.Tapsell. (May 2020). *Mitigating Disinformation in Southeast Asian Elections: Lessons from Indonesia, Philippines and Thailand*. NATO StratCom Centre of Excellence. Retrieved from www.stratcomcoe.org/mitigating-disinformation-southeast-asian-elections (Accessed on 12 October 2020).

186 OSCE. (2017). Joint Declaration on Freedom of Expression and 'Fake News'. *op. cit.*
UN Human Rights Council. *Disease Pandemics and the Freedom of Opinion and Expression*. A/HRC/44/49 (23 April 2020), para. 44.

187 2017 Joint Declaration on Fake News, para. 2(c) Retrieved from: https://www.osce.org/fom/302796 and (d); UN Human Rights Council, *Disease Pandemics and the Freedom of Opinion and Expression*, A/HRC/44/49 (23 April 2020), para. 44.

188 Organization for Security and Cooperation in Europe (OSCE). (30 April 2020). *Joint Declaration on Freedom of Expression and Elections in the Digital Age*. Para. 1(c)(i).

Chapter 3:
Platform Design and Reliability of Information

The COVID-19 pandemic has demonstrated the need to reverse the escalation of sensational content and rumor by promoting reliable news and information in a structured manner. Mechanisms and policies for promoting authenticity, reliability and findability of content are yet to be determined, based on established criteria.

Contents

INTRODUCTION

The International Declaration on Information and Democracy[189] states that **online service providers 'shall implement mechanisms that favor visibility of reliable information.** Such mechanisms shall be based on criteria of transparency, editorial independence, use of verification methods and compliance with professional norms. The integrity, authenticity and traceability of ideas and information shall be promoted, so that their origin and mode of production and dissemination are known. It shall not be a violation of political, ideological and religious neutrality to favor reliable information'.

This chapter will explore how the design of digital platforms' products could follow quality and safety standards, and how the design could contribute to elevate reliable information, add friction to exchanges to slow down the spread of disinformation, and contribute to breaking the content bubble of users.

While fact-checking is needed and should be expanded, **this chapter will highlight complementary options, including some ideas that might seem out-of-the-box or still at an exploratory stage.**

189 The International Declaration on Information & Democracy, adopted in November 2018 by the Commission on Information and Democracy, sets out the fundamental principles and guarantees of the global information and communication space.

1. PUBLIC STANDARDS FOR QUALITY AND SAFETY IN PLATFORM DESIGN[190]

1.1. A STATUTORY BUILDING CODE FOR DIGITAL PLATFORMS

Social platforms are no longer a mere microcosm of private users, they are spaces where our culture, economy and public discourse unfold. By the very nature of their prolific scale, utility and impact on society, social platforms have become affected with the public interest. Although social platforms are often publicly framed as 'free services',[191] **at the core of any social platform is a technical construction resulting from deliberate choices in design, architecture and engineering**. Services and engineering are often regulated through very different legal frameworks, norms and principles. In other technological sectors, certain aspects of design or engineering are not primarily governed by terms and conditions. This is because there is a paramount public interest in safety and quality that overrides any terms of private contract. In approaching the regulatory mix for social platforms, it is important to recognize a basic practical reality: **these platforms are engineered constructions as much as they are services.**[192]

Regulation of other technological fields focuses on maintaining **a common set of technical standards that inform engineers about the public's expectations for safety and quality.** These standards are often highly technical, written by domain experts working for regulators, and **mandate specific results from pre-defined tests that products must pass before being released.** In industry, **these standards are used by engineers to guide their work to surpass a desired threshold of safety and quality.** However, no such 'building code' exists for digital platforms, resulting in few substantive requirements for risk mitigation or safe design. In the course of ordinary software development, there is rarely any interaction with anyone charged with the protection of the public interest.

RECOMMENDATIONS TO **STATES**

> Explore and develop a new regulatory focus on digital architecture and software engineering in the regulation of online service providers.

> Collaborate with technical experts to design digital building codes for social platforms and other digital commons.

190 Proposals in section 1 by Christopher Wylie, Cambridge Analytica whistleblower and member of the Steering Committee of the working group on infodemics.
191 See, generally, the testimony of Mark Zuckerberg to the United States Committee on the Judiciary, 10 April 2018.
192 When looking at norms within the field of software development, it is clear that the profession understands the fundamental nature of its work as building. From the very job titles of 'engineer' and 'architect' to basic software concepts like 'building blocks', 'bridging', 'gateways', 'factories', 'ports', 'encapsulation', 'facades', 'filters' and 'wrappers', there is a common language of architecture and engineering that proliferates in the software profession.

> Develop specific and quantifiable tests[193] or technical standards for safety and quality that digital products must pass at their quality assurance (QA) stage of development, prior to public release.

1.1.a RE-ORIENTATION OF SOFTWARE ENGINEERING TOWARDS HARM PREVENTION

Building codes are oriented towards risk mitigation and the prevention of harm, where the precautionary principle is paramount. In these cases, the public interest is safeguarded not only by regulatory standards, but also inspections by technically competent authorities prior to public release. In this relationship, the engineer is not lawfully entitled to build at will. Rather, their work must be inspected by a third party. **In the case of civil engineering, there are no private 'terms and conditions' that can override the public's presumption of safety.**

In the same way that fire safety tests are conducted prior to a building being opened to the public, such a 'digital building code' would also result in a shift towards prevention of harm through testing prior to release to the public.

RECOMMENDATIONS TO **STATES**

> **Explore requiring minimum safety and quality standards centered on the precautionary principle.** This would help shift the perspective of designers, developers and engineers towards pre-emptively considering risks at the early stages of product development. It would also empower software professionals to take a more proactive role in resisting problematic commercial pressures to implement what are collectively termed 'dark pattern designs'[194] where such design undermines user choice or privacy.

1.1.b AGENCY BY DESIGN

The prolific reliance of platforms on catch-all terms and conditions is premised on a rational choice model of disclosure decision-making, or 'privacy pragmatism', whereby users are presumed to balance their privacy preferences with their desire to use a platform.[195] However, there is a growing corpus of behavioral research that disputes this assumption of user rationality.[196] In his book *Mindf*ck*,[197] data scientist and Cambridge Analytica whistleblower Christopher Wylie argues that platforms are not always

193 A falsifiable test is a test where there is a clear yes/no answer. For example: 'When applied to a test set of diverse users, are there any observed statistical effects arising from the decisions of the trained algorithm that could materially disadvantage a user on the basis of a protected category, such as gender or race?'

194 Dark patterns are duplicitous user experience (UX) designs that are implemented to confuse, manipulate, distract or frustrate a user into behaving on a website in a way that they otherwise would not behave. According to darkpatterns.org, a site that documents the issue, some examples of common tactics include: 'Roach Motel' (where opting into a feature is easy, but opting out is difficult or obscured), 'Privacy Zuckering' (where a site tricks a user into unintentionally sharing private information), 'Misdirection' (where an interface distracts a user from realising an unwanted functionality), and 'Confirmshaming' (where the wording for declining an option is phrased to shame the user into compliance).

195 Draper, N. (2017). From Privacy Pragmatist to Privacy Resigned: Challenging Narratives of Rational Choice in Digital Privacy Debates. *Policy & Internet.* Volume 9, Issue 2.

196 Keys, D. & Schwartz, B. (2007). 'Leaky' Rationality: How Research on Behavioral Decision Making Challenges Normative Standards of Rationality. *Perspectives on Psychological Science.* Volume 2, Issue 2.

197 Wylie, C. (2019). *Mindf*ck – Cambridge Analytica and the Plot to Break America.* New York: Random House.

honest brokers in their reliance on standard form terms: 'However, too often privacy is eviscerated through the bare performance of clicking "accept" to an indecipherable set of terms and conditions. This consent-washing has continually allowed large tech platforms to defend their manipulative practices through the disingenuous language of "consumer choice". This positions our frame of thinking away from the design—and the designers—of these flawed architectures, and toward an unhelpful focus on the activity of a user who does not have understanding or control over the system's design.'

Through reliance on confusion, deception or manipulation to achieve the bare performance of user acceptance, dark patterns are fundamentally incongruent with any substantive requirement of informed consent or giving effect to genuine consumer choice online, considers Wylie.

RECOMMENDATIONS
TO **STATES**

> Consider restricting or banning dark pattern design on digital platforms.

> **Consider creating a new legal principle of 'agency by design'.**[198] Agency by design would require design to be proactively choice-enhancing. Statutory tests could be created to give life to this principle. Design could be analyzed through the prisms of reasonable expectation (e.g., whether an average customer would reasonably expect and understand the holistic outcomes of a given feature, app or option) or proportionality of effects (e.g., whether the effects of a given feature, app or option are proportionate to the user's legitimate interests, including a user's right to agency and privacy).

1.1.c STATUTORY LIMITATIONS ON ALGORITHMIC UNDUE INFLUENCE

When self-learning algorithms are coupled with vast amounts of personal data, platforms gain significant informational asymmetries. These asymmetries have led to gross power imbalance in the user-platform relationship. This imbalance is not simply present in targeted advertising, but in the entire informational ecosystem, such as newsfeeds, timelines and groups. As a result, users can be influenced into making decisions that they may not have taken otherwise.

Although there is a diverse set of risk scenarios, a common feature of algorithmic harms is that they exert an unfair influence on a person's ability to make informed choices. This is the case with misinformation (e.g., anti-vaccine conspiracies) and in some instances of algorithmic bias (e.g., algorithms obscuring property adverts from ethnic minorities). **The harm that results for the user in these scenarios stems from a denial of agency, i.e., from depriving someone of their ability to exercise freedom of thought. Although the user is still making choices, their decisions ultimately flow from understandings that may suffer from detrimental misconceptions that were promoted by a platform's algorithms.** This could result in individual choices that would not occur but for the duplicitous intervention of an algorithm to amplify or withhold select information on the basis of engagement metrics created by a platform's design or engineering choices.

198 This principle takes inspiration from the Canadian and European principle of 'privacy by design', which re-positioned privacy into a practical design requirement. See: Cavoukian, Ann. (August 2009). Privacy by Design: The 7 Foundational Principles. Office of Information and Privacy, Commissioner of Ontario. August 2009.

A user risks becoming a vulnerable party in cases where they make detrimental choices that they otherwise would not have made were it not for the intervention of an algorithm. In such cases, users may experience a manifest disadvantage from ensuing actions that can be traced back to the dominating effects of a social platform's algorithms. **In other areas of law, there is an established doctrine of undue influence.** Although the specifics vary across legal systems, **undue influence is broadly where one party's influence upon the other is so disproportionate, powerful or malign that it calls into question whether the decisions or consent of the other party are truly valid.** The disproportionate effect that scaled artificial intelligence can have on the exercise of individual agency is analogous and poses a threat to the exercise of individual agency.

Already, the deleterious impact of algorithmic amplification of COVID-19 misinformation has been seen, and there are documented cases where individuals took serious risks to themselves and others as a result of deceptive conspiracy theories presented to them on social platforms. The law should view these individuals as victims who relied on a hazardously engineered platform that exposed them to manipulative information that led to serious harm.[199]

RECOMMENDATIONS TO **STATES**

> **Consider creating a doctrine of 'algorithmic undue influence'. Known and familiar legal constructs can be adapted to address newfound harms arising from the emergence of artificial intelligence.**

> Form working groups of machine learning engineers and lawyers to further explore and adapt this concept into domestic legal systems.

1.1.d SOFTWARE MALPRACTICE IN CASES OF MANIFESTLY NEGLIGENT DESIGN

Harms on social platforms are often the result of a lack of due care or consideration for how architectural choices may expose the public to risk. This is despite the fact that there are widely known, documented and foreseeable risks of poorly constructed social architectures lacking fail-safe mechanisms.[200] There should be legal recourse for these very real harms.

RECOMMENDATIONS TO **STATES**

> **Explore creating legal liability for 'software malpractice'** in cases where it can be shown that the engineering or design choices of a platform's architecture created manifest and foreseeable risks of harm. Reframing liability in terms of 'software malpractice' also will allow states to address manifest harms online without necessarily wading into complex debates about free speech or liability for

199 There is precedent for this in other areas of consumer law where, for example, deceptive advertisers are liable for harms to consumers who detrimentally rely upon deceptive information.

200 For example: the incitement of hate, racial or religious violence, public health misinformation, state-backed cyber propaganda and political disinformation.

user generated content.[201] Rather, **liability for software malpractice would focus solely on whether or not specific design standards were followed and whether appropriate risk mitigation features were in place at the time.**[202] This would allow regulators to pre-emptively outline and address tangible online harms, while also balancing the needs of industry for both regulatory clarity and the avoidance of spiraling theoretical liability for the actions of every user.

1.1.e DE-SEGREGATE THE DIGITAL COMMONS

Social platforms often describe themselves in the language of 'communities'. However, as they structure the global information and communication space, platforms also tacitly assume a role akin to town planners. However, they do so without any corresponding duties to public safety. Some of the most insidious harms documented from social platforms have resulted from the algorithmic herding of users into homogenous clusters whereby users engage only with filtered content and communicate with other users in those same clusters. In this process, newsfeed algorithms and recommendation engines 'curate' information that reinforce biases, which has been attributed to increasingly fractured public discourse.[203] There is also evidence that social platform algorithms have contributed to radicalization and extremism online.[204]

Although innocuously termed by industry as 'filter bubbles', this effect could also be described as a form of segregation. In his book, Wylie argues that: **'What we're seeing is a cognitive segregation, where people exist in their own informational ghettos. We are seeing the segregation of our realities. If Facebook is a "community," it is a gated one.** Shared experience is the fundamental basis for solidarity among citizens in a modern pluralistic democracy, and the story of the civil rights movement is, in part, the story of being able to share space together: being in the same part of the movie theatre or using the same water fountain or bathroom.'[205] **Creating shared digital commons without segregation should be an objective of digital regulation.** In many countries, it is unlawful for architects or planners to develop racially or religiously segregated buildings or neighborhoods. The substantive harm of cognitive segregation online is analogous to civil segregation and ghettoization.

RECOMMENDATIONS TO **STATES**

> Consider applying equivalent anti-segregation legal principles to the digital commons, including a ban on 'digital redlining', where platforms allow groups or advertisers to prevent particular racial or religious groups from accessing content.

> Create legal tests focused on the ultimate effects of platform design on racial inequities and substantive fairness, regardless of the original intent of design.

> Create specific standards and testing requirements for algorithmic bias.

201 For example, the so-called 'Section 230' debate in the United States.
202 To use a physical analogy, in the event of criminal arson, an architect or building owner are not liable for the direct actions of a criminal arsonist. However, they would be liable if they did not put in the minimum number of fire exits per the building code and where people were harmed because it was difficult to escape.
203 Sîrbu, A., D. Pedreschi, F. Giannotti, & J. Kertész. (2019). Algorithmic Bias Amplifies Opinion Fragmentation and Polarization: A Bounded Confidence Model. *PLOS One*. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213246 (Accessed on October 21, 2020).
204 Sonnemaker, T. (2020). Facebook Reportedly Had Evidence That its Algorithms Were Dividing People, But Top Executives Killed or Weakened Proposed Solutions. *Business Insider.* Retrieved from https://www.businessinsider.fr/us/facebook-knew-algorithms-divided-users-execs-killed-fixes-report-2020-5# (Accessed on October 21, 2020).
205 Wylie, C. *op. cit.*

1.2. REGULATORY AUDITS FOR SAFETY AND QUALITY ASSURANCE

In other fields of engineering, **compliance with safety standards can often be *proven or disproven* to a regulator with observable test cases because requirements are highly articulated and falsifiable.** In software engineering, it is already normal practice to build in quality assurance (QA) tests to check the performance of code before it is released. A digital building code would build upon this existing QA practice and require that a developer additionally use a set of externally defined test cases during the QA phase of software development. This would not only require better risk mitigation practices at the pre-release stage of development, it would also provide platforms greater legal certainty that if the tests are passed, a product would be safe enough for public release. Regulation need not be burdensome for platforms. Well-articulated test cases outlined in technical regulation could help achieve a platform's obligations with clarity and confidence. There are a plethora of issues where minimum standards could be enumerated.

RECOMMENDATIONS TO **STATES**

> Within a digital building code, specifically **articulate the risks that need to be mitigated**, and how online service providers can ensure minimum compliance through pre-set tests that can be implemented in QA stages of software development.

> Consider requiring online service providers to deploy recommendation algorithms in 'virtual sandboxes'[206] to undergo safety and quality testing, prior to releasing them into the public. These tests could explore an algorithm's potential biases towards misinformation or amplification of racial hatred.

> Consider requiring that platforms subject new features to 'abusability testing'[207] as an addition to the already standard practice of usability testing.

> To ensure clarity, the digital building code should outline specifically what sandbox environments would need to contain, what quantifiable tests would need to be conducted, how they should be performed and documented, and what is the minimum threshold for passing the test.

206 Virtual sandboxes are used in software development to test the functioning of code and the behaviours of an algorithm in a simulated environment. This allows a developer to identify flaws or vulnerabilities in an isolated environment without risking the integrity of a live network or site. As a note, there are some examples of where national regulators have started to work alongside industry with sandbox environments, such as the UK Information Commissioner's Office sandbox programme.

207 This concept would extend the common software development practice of usability testing, where user behaviour on an frontend interface is studied by UX developers to improve functionality, and instead would focus on identifying ways that a feature could be *abused* by users. In the context of social platforms, this would include proactive attempts at manipulating or undermining a site's features or algorithm's behaviour to identify vulnerabilities, risks and unintended consequences.

1.3. BOTTOM-UP REGULATION WITH PROFESSIONAL CODES OF CONDUCT FOR SOFTWARE ENGINEERS

Much of the debate around social platforms centers on the regulation of companies, but it should be remembered that the underlying software of platforms is designed and made by professional software engineers.

In his book, Wylie argues that: 'Out of all the possible types of regulation, **a statutory code for software engineers is probably what would prevent the most harm**, as it would force the builders themselves to consider their work before anything is released to the public and not shirk moral responsibility by simply following orders. Technology often reflects an embodiment of our values, so instilling a culture of ethics is vital if we as a society are to increasingly depend on the creations of software engineers. If held properly accountable, software engineers could become our best line of defence against the future abuses of technology. And, as software engineers, we should all aspire to earn the public's trust in our work as we build the new architectures of our societies.'[208]

In other sectors, design liability establishes the standard of care that must be applied by the designer or architect. Many legal systems require that an engineer or architect design their constructions with reasonable skill and care. This liability may be reduced in cases where designs are technologies at the precipice of the state of art. However, there are precedents in law where courts have found that even if a design or technology is 'beyond the frontiers of professional knowledge at that time it [is] still incumbent on [the designer] to exercise a very high degree of care'.[209] The goal in creating a code of ethics for software engineers would **add a much-needed 'bottom-up' effect to regulation.** Too often engineers do not feel empowered to challenge their employer to incorporate more ethical practice into software design, and such a professional code would catalyze healthy friction within platform companies.

RECOMMENDATIONS TO **STATES**

> Consider creating **an enforceable code of ethics for software engineers, backed by a statutory professional body.** The code should outline specific duties, including a duty to refuse and a duty to report negligent practices where there are manifest risks to the public. The code should not only outline professional penalties for software malpractice, but should also embolden professional engineers to take a more proactive role in considering the public interest at work.

208 Wylie, C., *op. cit.*
209 See, Independent Broadcasting Authority v EMI Electronics and BICC Construction Ltd [1980] 14 BLR 1

1.4. ENFORCEMENT THROUGH A PUBLIC DIGITAL STANDARDS ENFORCEMENT AGENCY

There is a common view that the 'law cannot keep up with technology'. However, simply because a technology is complex or fast-moving does not mean there are no ways of effectively creating rules to protect the public interest. Wylie notes that: 'Tech companies should not be allowed to move fast and break things. Roads have speed limits for a reason: to slow things down for the safety of people. A pharmaceutical lab or an aerospace company cannot bring new innovations to market without first passing safety and efficacy standards, so why should digital systems be released without any scrutiny?'

In other technical fields responsibility is delegated to technically competent regulators specifically because it is difficult for lay parliamentarians to keep up with those innovations. In many other sectors, including technically complex areas such as pharmaceutical research, telecoms infrastructure or nuclear engineering, governments have created statutory bodies empowered to create technical regulations, inspect the work or products of companies, and issue enforcement orders where minimum standards are not met.

Rules without enforcement are impotent, and any regulatory framework must consider not only what rules should be created, but who will ensure that those rules are respected. Although some jurisdictions, such as member states in the European Union, have created regulatory bodies for data protection, this framing around data protection is self-limiting as it does not address the wider array of risks arising from digital platforms beyond privacy.

RECOMMENDATIONS TO **STATES**

> **Consider creating a 'Digital Standards Enforcement Agency'** with wider regulatory competence to address the digital sector as a whole. Similar to technical regulators in other industries, such an agency could be given the powers to:
- **Identify and define harms and safety risks in the digital context;**
- **Draft, maintain and update minimum design and architecture standards;**
- **Enforce professional standards in software engineering;**
- Investigate user complaints and inspect technology firms;
- Issue enforcement orders for non-compliance; and
- Prosecute persistently non-compliant offenders.

> The Forum on Information and Democracy could launch a feasibility study on the implementation of this agency.

2. ELEVATE RELIABLE INFORMATION

2.1. PROHIBIT CONFLICTS OF INTEREST

Online service providers must be subject to an obligation of neutrality in relation to their own interests. They must be required to represent reality with sincerity, that is to say they should not limit its representation to the content, goods or services they have an interest in. Moreover, they should not favor or disfavor content, goods and services, for this reason.

RECOMMENDATIONS TO **STATES**

> **Consider imposing on online service providers a set of obligations of neutrality with regard to their own interests** (see above). Obligations should include:

- When online service providers highlight a piece of content, goods or a service, they must be required to clearly indicate contractual or commercial interests they have with them. They should also reference other content, goods or services.
- Online service providers should not be able to promote commercial interests to which they are linked without clear transparency.
- The promotion of content, goods or services to which they are linked (contractually or commercially) must not prevent, limit or even influence the referencing of other content, goods or services.
- Any situation of conflict of interest, i.e., any breach of this obligation of neutrality and transparency on their interests, must be subject to sanctions.

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Implement their transparency obligations to enable the independent regulatory authority to verify compliance with the neutrality obligation** (see Chapter 1 on transparency requirements for online service providers).

2.2. IDENTIFY RELIABLE SOURCES OF INFORMATION

A structural approach to furthering the reliability of information can be followed by means of **promoting information that has been produced in adherence with an internationally accepted set of best-practice guidelines and ethical norms. Public interest journalism that meet professional and ethical standards could be considered as reliable information.** Such an approach would help consumers and citizens, advertisers, distributors and regulators to identify and reward trustworthy content.[210]

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Cooperate with existing efforts to create instruments for enabling trustworthy news and information.** These instruments must result in machine-readable signals to inform both human and algorithmic decision-making in content distribution and consumption.[211]

RECOMMENDATIONS TO **JOURNALISTIC COMMUNITIES**

> Participate in establishing unified benchmarks based on internationally accepted best-practices and ethical norms.

> Standards must be data-driven, that is to say that the standard clauses must be machine-readable to inform both human and algorithmic decision-making in the distribution and consumption of content.

> These technical standards should meet two safeguards:[212]

- **(1) Legitimacy:** Providers of public interest journalism can be identified through voluntary, self-regulatory standards, which are transparently sourced, governed and enforced. These standards must be based on internationally accepted best-practices and ethical norms to serve as legitimate criteria for the duty of due prominence, defined in part 2.3 of this chapter. The application of these technical standards must be attributed and disclosed by and to all parties involved.
- **(2) Neutrality:** Technical standards for the application should be publicly available to any interested party (conforming entities and service providers, consumers or users, public authorities, etc.). As regards conformity assessment, standards should be neutral so that they do not prescribe any particular way of conformity assessment (i.e. self-declaration, verification by the purchaser or testing/certification by an independent body in accordance with existing accreditation schemes).

210 Reporters Without borders (RSF) contribution to the working group on infodemics.
211 Steenfadt, O. (2020). Sustaining Journalism During COVID-19, How the EU Can Turn Digital Platform Regulation Into a Tool for Democracy. Friedrich-Ebert-Stiftung.
Retrieved from http://library.fes.de/pdf-files/bueros/budapest/16406.pdf (Accessed on 2 October 2020).
212 Steenfadt, O., E. Mazzoli, & S. Luca. (2020). Ensuring The Visibility and Sustainability of Reliable, Accurate Information in the Online Sphere Through Voluntary Standards – A Co-Regulatory Approach (Accessed on 2 October 2020).
It define two main safeguards as formulated below.

Such an approach has already been initiated by the Journalism Trust Initiative (JTI),[213] a collaborative process of standardization designed to encourage respect for journalistic ethics and methods and reinforce the right to information by promoting online content produced in accordance with these principles. The process was launched under the aegis of the European Committee for Standardization (CEN) to draft a 'workshop agreement'. More than 120 entities collaborated in the drafting, including news agencies, media unions, consumer groups and online service providers. The JTI European 'Standard' was officially issued on 19 December 2019 in the form of a 'CEN workshop agreement'.[214]

Once a standard text has been adopted and published, it can be used by different stakeholders and in different ways:

- online service providers can use the data to promote conforming sources;
- advertisers can use it to align their spending, and thus enhance brand safety;
- regulators and governments can use it as a normative mechanism to allocate subsidies (including Corona-related aid packages) to eligible media outlets;
- media development actors and donors can use it as a tool to evaluate partners and beneficiaries, but also for capacity building in the field of trust, governance and media management systems.

The next step of the implementation of the standard encompasses two levels:

- **Self reporting:** outlets can voluntary and proactively disclose their identity and inner workings in order to enhance trust in their own best interest. Even if not audited, this would already manifest a major achievement in terms of transparency, promoting media and information literacy.
- **Conformity assessment:** an audit of the self-reported disclosures by a trusted third party to gain recognition.

2.3. INCLUDE RELIABILITY IN ALGORITHMS

In the same way as 'must-carry' rules require cable operators to carry local broadcast television stations so they do not lose market share, lawmakers could create an obligation for online service providers to implement mechanisms aimed at highlighting information sources that comply with standardized professional and ethical self-regulation standards ('must-be-found'), to give them preferential treatment in terms of content prioritization, and more particularly promotion and visibility in news feeds and research results.[215]

RECOMMENDATIONS
TO **STATES**

> **Consider initiating a co-regulatory framework for the promotion of public interest journalism which would entail respective legal obligations for online service providers**, and which should be stipulated by law, while the actual specifications are to be determined in a self-regulatory way with reference to these technical standards in the law (see part 1 of this chapter).

213 The process was launched by Reporters Without Borders (RSF) under the aegis of the European Committee for Standardization (CEN), with help from French CEN member AFNOR and German CEN member DIN, and in partnership with Agence France-Presse (AFP) and the European Broadcasting Union (EBU), to draft a 'workshop agreement'. More than 120 entities collaborated in the drafting , including news agencies (such as the Associated Press, DPA and EFE), broadcast media (such as the BBC, RTL and France TV), media unions (journalists' federations in Taiwan and South Korea), consumer groups (the European Consumer Organisation) and tech groups (such as the World Wide Web Consortium). Google and Facebook participated, as did many regulatory bodies from various European countries.

214 The CEN Workshop on Journalism Trust Initiative publishes its CWA 17493. (2019). Retrieved from https://www.cen.eu/News/Workshops/Pages/WS-2019-018.aspx (Accessed on 2 October 2020).

215 This co-regulatory framework was proposed by Steenfadt et al. (2020). *op. cit*., where they formulated the recommendations below for lawmakers.

> Adherence to the technical standards by media outlets must be voluntary, should not be discriminated against, and should be possible without undue limitation of competition.

> The co-regulatory framework should concern platforms that are offering software-based services that actively influence the flow of online information between providers and consumers, by structuring, curating and recommending content. This category potentially includes, but is not limited to, news content aggregators, search engines and social media networks.

Legal obligations of the co-regulatory framework should include:

> **(1) Due Prominence**: Online service providers that actively structure or curate content should be obliged to ensure due prominence of public interest journalism on their services. Services that cater to special interests may be exempted from this obligation. Appropriate prominence measures include the use of technical standards as stipulated in part 1 of this chapter—standards established in a participatory and transparent manner in order to identify entities operating according to the highest internationally recognized professional norms, so as to produce reliable and accurate information.

> **(2) No-harm Principle**: Appropriate measures as per this provision should not discriminate on the basis of content or viewpoint. Online service providers should not treat non-compliance with or non-usage of technical standards to identify reliable sources of information as a reason to exclude, downrank, demote or otherwise actively affect the visibility or monetization of content in a negative way.

> **(3) Transparency**: In order to demonstrate compliance with their duty to ensure due prominence for public interest journalism on their services, online intermediaries should establish mandatory transparent mechanisms and metrics of indexation regarding the discoverability and visibility in search ranks, news feeds and products, including the provision of data and information on prioritization, personalization, and recommendation algorithms, audits and complaints, in an accountable manner.

> **(4) Enforcement and Evaluation**: National Regulatory Authorities should monitor and assess whether appropriate measures adopted by online service providers under the due prominence obligation are sufficient to contribute to pluralism and diversity in their respective national markets.

- National Regulatory Authorities should rely on self-regulatory and co-regulatory mechanisms such as Social Media Councils or E-courts to ensure a broad, open and transparent participation of all stakeholders in the assessment process.
- The regulator is assigned the task of assessing the effectiveness and impact of the duty of due prominence under (1), notably by compiling information from National Regulatory Authorities in a public report twice a year.

> **(5) Consumer Choice**: Individual users must always have a clear and easily accessible choice to opt out of the appropriate measures designed to ensure due prominence to public interest journalism.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Cooperate with existing efforts to create instruments for enabling trustworthy news and information on their services.

> Measures taken by platforms to promote authoritative news and information sources should be grounded on a standard based on internationally accepted best-practices and ethical norms.

> When applying standards to their algorithmic mechanisms, platforms should be transparent about how they apply to their recommendation algorithms. This should be explained proactively in language easily understandable by all those affected by it.

> Track the effectiveness of implementing such technical standards in their algorithmic recommendation mechanisms in order to avoid unintended issues that might arise from their use.

3. CREATE FRICTION

Recent examples of viral COVID-19 hoaxes have shown the challenges faced by online service providers in countering the spread of items of content.[216] **Online service providers are always improving their interfaces to suppress any friction, in order to optimize users' experience and maximize time spent on their platform. But in order to fight disinformation, online service providers are now considering ways of creating friction to inhibit the spread of potentially unreliable content, so as to offer more contextual information to users and empower them.**[217]

> *"From a surveillance capitalist's standpoint, any friction means less behavioral data, ergo, less profits."*
>
> **Gelo Gonzales, technology editor at Rappler.**[218]

216 In May 2020, despite Facebook and YouTube's efforts, a single upload of the 'Plandemic' video got 7.1 million views before it was removed. See, Newton, C. (2020). A Plandemic Sequel Goes Viral. *The Verge*. Retrieved from https://www.getrevue.co/profile/caseynewton/issues/a-plandemic-sequel-goes-viral-267019 (Accessed on October 10, 2020).
In July 2020, the promotion of a mysterious unproven COVID cure by the group America's Frontline Doctors reached 20 million views on Facebook in less than 24 hours. See Newton, C. (2020). New Ideas for Fighting COVID-19 Misinformation. *The Verge*. Retrieved from https://www.getrevue.co/profile/caseynewton/issues/new-ideas-for-fighting-covid-19-misinformation-272134 (Accessed on 9 October 2020).

217 In October 2020, Twitter announced additional steps to 'slow the way information flows on its network, even changing some of its most basic features'.
Conger, K. (2020). Twitter Will Turn Off Some Features to Fight Election Misinformation. *New York Times*.
Retrieved from https://www.nytimes.com/2020/10/09/technology/twitter-election-ban-features.html (Accessed on October 10, 2020).

218 Gonzales, G. , *op. cit.*

3.1. CONTEXTUALIZE AND LABEL

3.1.a CONTEXTUALIZE

Users should be given additional context to help them process information. Online service providers tend to present all information the same way — despite varied sources, relationships, history, or purpose mediating relationships between content producers and content consumers.[219]

For example, Reuters Institute for the Study of Journalism recently found that 59% of posts on Twitter which were rated as false by fact-checkers remained up without warnings, whereas on Facebook 24% of false-rated content in a particular sample remained up without warning labels.[220]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Automatically propose a plurality of information sources or dedicated information centers for posts on topics related to electoral processes and health.**

> **When applicable, provide contextual information about the user posting, including details such as location, relationship to the reader, duration on the platform, institutional affiliations, and verified expertise in key areas of public interest.[221]**

> Provide contextual information about the source of third-party content, such as content publication dates, host sites, details inferred from top-level and second-level domains, and on-platform details, as well as information from verification programs about the third party, such as whether the URL is frequently fact-checked on a platform, or whether off-platform information is drawn from groups with strong verification processes, such as Wikipedia.[222]

> **Provide contextual information about how the content has been shared or promoted on the online service provider's network.**

> Major social media platforms should compensate any independent entities whose work is used to help improve quality or provide contextual information, including fact-checking organizations, Wikipedia, and independent media groups—even if licensing allows free use.[223]

> **When independent fact checkers determine that a piece of content is disinformation, the platforms should show a correction to each and every user exposed to it—that means anyone who viewed, interacted with, or shared it. This can cut belief in false and misleading information by nearly half.[224]**

219 Simpson, E., & A. Conner. (2020). Fighting Coronavirus Misinformation and Disinformation. Center For American Progress. Retrieved from https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation/ (Accessed on 8 October 2020).
It proposes recommendations to fight coronavirus misinformation and disinformation.
220 Brennen, S. et al. (2020). Types, Sources, and Claims of COVID-19 Misinformation. Reuters Institute. Retrieved from https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation (Accessed on October 20, 2020).
221 Simpson, E. & A. Conner. *op. cit.*
222 Simpson, E. & A. Conner. *op. cit.*
223 Simpson, E. & A. Conner. *op. cit.*
224 Avaaz contribution to the working group on infodemics.

3.1.b REQUIRE DISCLOSURE LABELING FOR TARGETED ATTRIBUTES

Ordinary users are often unaware of how much content presented to them is targeted or curated as a result of algorithmic profiling based on their personal data. This is important context that, if made more salient, could have an impact on how a user interacts with an advertisement or newsfeed item. Some platforms have begun offering users basic information about why they are being shown advertisements, but too often this information is not detailed, offering vague explanations, such as that the advertiser wanted to reach users 'similar to their customers'. Instead, users should be entitled to view a complete and detailed set of attributes that recommendation engines or targeting algorithms are using to decide to show them content. This could also involve requiring platforms to show the specific 'feature importance' scores of an algorithm, to highlight not only what attributes about the user the algorithm is using, but also how important those features are in that context.[225]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Disclose a complete and detailed set of attributes that recommendation engines or targeting algorithms are using to decide to show users' content.**

3.1.C LABEL STATE-CONTROLLED MEDIA

In an attempt to combat state propaganda, YouTube,[226] Facebook[227] and more recently Twitter[228] have started to label media that are 'state-funded', 'state-controlled' and 'state-affiliated' respectively. However, this approach presents many flaws, including inconsistency of the criteria, their arbitrary application between media outlets, and a lack of transparency over the governance and implementation of the criteria. While definitions differ from one online service provider to another, they often exclude outlets controlled by political parties (in power or opposition), and groups aligned with political, economic and religious interests, and neglect sub-national, or even local media, which can advance a political agenda and impact local elections just as violently. They also exclude brand inconsistencies that could arise between different regional branches of the same media.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Agree on a common definition throughout the industry. **The definition should not be subjective but based on tangible criteria, publicly available and built in cooperation with the journalistic community and civil society.**

> **Make a distinction between 'state-controlled' media and public media.**

225 In statistical modelling, feature importance ranking shows the relevance of each input variable used by an algorithm, including the degree of positive or negative impact that variable has on the prediction or classification.
226 In February 2018, YouTube added an information label besides individual video posts of 'state-funded' media, including public media. See, Samek, G. (2018). Greater Transparency for Users Around News Broadcasters. YouTube Official Blog. Retrieved from https://blog.youtube/news-and-events/greater-transparency-for-users-around?m=1 (Accessed on 2 October 2020).
227 In 2019, Facebook announced its new policy to label 'state-controlled' media and this was implemented in June 2020. Labels can be found in the transparency sections of profile pages, and since recently, in the United States, posts appearing in news feeds are individually labeled.
228 In August 2020, Twitter introduced 'state-affiliated' labels on profile pages and posts.

> **Be transparent** about:
  - the criteria used to define 'state-controlled' media;
  - the governance and implementation of the criteria;
  - the effects of being labeled on the visibility of media outlets and their ability to advertise;
  - the list of media labeled.

> **Notify media accounts when they are identified and labeled as such.**

> Create appeal processes and mechanisms for redress, and ensure due process.

> Track data to assess the impact and effectiveness of these labels on educating audiences and reducing disinformation campaigns. Global results should be published in online service providers' periodic transparency report. More in-depth data could be provided to vetted academics to conduct further research (see Chapter 1).[229]

3.2. BACK-END FRICTION

3.2.1. COOLING-OFF PERIODS

One approach[230] to reducing the impact of content bubbles (see part 4 of this chapter) could be to **require 'cooling-off periods', where after a set number of impressions from a common advertiser, a platform's recommendation engine would be required to deliberately switch to displaying different content.** This could go beyond simply targeted advertising, and require newsfeed or timeline algorithms to behave similarly in their curation of organic content. This would be achievable through using statistical tests of content similarity, which is already a well-developed area of machine learning.[231]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Implement cooling-off periods for targeted advertising and organic content, so that beyond a certain threshold of impressions a platform's recommendation engine would be required to deliberately switch to displaying different content.**

> **Limit the number of times similar types of organic content or advertisements from the same advertisers is seen by one user.**

3.2.2. CIRCUIT-BREAKER

When content becomes viral, it reaches a large audience in the network before online service providers can evaluate it and enforce their terms of service.

229 Radsch, C. (2020). Tech Platforms Struggle to Label State-Controlled Media. Committee to Protect Journalists. Retrieved from https://cpj.org/2020/08/tech-platforms-struggle-to-label-state-controlled-media/ (Accessed on 2 October 2020).
230 This approach was proposed by Christopher Wylie, member of the Steering Committee of this working group in infodemics.
231 For example, in the area of natural language processing (NLP), there is a wide array of techniques that can be used to achieve this objective.

Researchers Ellen P. Goodman and Karen Kornbluh compare viral content online to high-frequency trading. On financial markets, circuit breakers have been used to prevent the panic associated with market volatility and excessive speculative gains or losses. **A circuit breaker is triggered when stock prices drop by a certain percentage from the previous day's closing price. It stops trading for a period of time, so that investors have the opportunity to understand what is happening in the market and act accordingly.[232] A similar mechanism could be implemented by online service providers to slow down the spread of potentially harmful viral content until it can be reviewed by human moderators.**

RECOMMENDATIONS TO **STATES**

> **Consider creating a duty for online service providers to disrupt traffic at a certain threshold of reach. Human review would then be required to assess the communication for compliance with applicable laws and terms of service/ community guidelines and standards.**[233]

RECOMMENDATIONS TO **SERVICE PROVIDERS**

Implement circuit breakers for viral content.

> When content reaches a certain threshold, online service providers **should trigger an internal viral circuit breaker that would temporarily prevent the content from algorithmic amplification in newsfeeds,** appearing in trending topics, or via other algorithmically aggregated and promoted avenues. Individual posting or message sharing could still occur. The algorithmic pause would allow the necessary time for a platform to review the content.[234]

> **Viral content should automatically be placed at the top of a queue for third-party fact-checking.**

> Online service providers should be transparent about how their circuit breakers operate. They should track the effectiveness of implementing such tools, and should partner with researchers to enable independent study of these interventions.

232 Goodman, E. & K. Kornbluh. (2020). Social Media Platforms Need to Flatten the Curve of Dangerous Misinformation. *Slate*. Retrieved from https://slate.com/technology/2020/08/facebook-twitter-youtube-misinformation-virality-speed-bump.html (Accessed on 8 October 2020).
233 Goodman, E. (2020). Digital Information Fidelity and Friction. Knight First Amendment Institute at Columbia University. Retrieved from https://knightcolumbia.org/content/digital-fidelity-and-friction (Accessed on 8 October 2020).
234 Simpson, E., & A. Conner. *op. cit.*

> In order to create a body of work to inform the development of a viral circuit breaker, online service providers should begin by using internal data from past user interactions and identified examples of viral misinformation in order to examine the spread of previous misinformation retroactively. That analysis should then be used to identify common patterns among viral disinformation so as to model the impact of potential interventions. Online service providers should rapidly and transparently collaborate, test, and identify reliable indicators of harmful posts to carefully hone such a detection system—opening this process to contribution from researchers, journalists, technologists, and civil society groups across the world. Trending posts that have reliable indicators of mis/disinformation should trigger rapid review by content moderation teams and get prioritization within fact-checking processes.[235]

4. BREAK THE CONTENT BUBBLE

4.1. IMPOSE A MANDATORY LEVEL OF NOISE

"The problem with some of these algorithms is that they become too good and too focused on particular user's attributes. It will give you only what the algorithm knows you will like and engage with," explains Christopher Wylie[236]. This gives rise to the idea of the need for **defocusing the algorithm by imposing a mandatory level of noise: a mandatory level of random content included in the algorithm.**

When an algorithm is developed, it is 'trained' on a set of data. Within training sets, there are two types of information: signal (helpful information) and noise (random or unhelpful information). Adding a mandatory level of noise in the training set of the algorithm would prevent any recommendation engine from becoming too focused or targeted, and could help disperse the focus of the information seen by the user and prevent filter bubbles being created.

The random content introduced in the algorithm could be accessed potentially from content published by entities that provide public interest journalism identified using a standard based on internationally accepted best-practices and ethical norms for the production of reliable information, as defined in part 2.2 of this chapter.

RECOMMENDATIONS
TO **STATES**

> **Consider developing technical standards for minimum noise in training sets used to train recommendation engine algorithms on social platforms.**

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Include a certain level of randomness in the training set of the relevant algorithms to prevent the creation of filter bubbles.**

235 Simpson, E., & A. Conner. *op. cit*
236 In interview with rapporteurs.

4.2. CONSIDER LIMITING MICRO-TARGETING

Some experts consider that banning or limiting micro-targeting could reduce the spread of disinformation and misinformation, while acknowledging that **micro-targeting is at the core of the business model of digital platforms. It is important also to keep in mind that there are positive uses of microtargeting, such as helping unregistered voters to register to vote. Such exploration around micro-targeting should begin with the difficult exercise of defining which micro-targeting should be limited, and with an understanding of potential unintended consequences.**

Nonetheless, micro-targeting can be dangerous not only because it invades the user's privacy but because the information provided is shaped for the targeted audience, notes the director of the Technology, Media, and Communications specialization at Columbia University School of International and Public Affairs, Anya Schiffrin.[237] The international human rights lawyer Susie Alegre warns that '**the intent to access our thoughts on an individual level without our knowledge and use that information to change our thoughts, emotional states, opinions and therefore our voting behavior without us realizing it must surely amount to an attempt to interfere with freedom of thought on a grand scale**.... Behavioral microtargeting is of particular concern because it aims to interfere with our thoughts for political gain and is a threat to the foundations of our democracies'.[238]

Alex Campbell, researcher at the Center for Global Security Research at Lawrence Livermore National Laboratory, considers that data privacy laws could offer 'an elegant arrow in the quiver of responses to online disinformation, intervening directly in the machinery of microtargeting essential to disinformation campaigns. Because they are adversary-agnostic, such laws protect against foreign and homegrown trolls alike, while avoiding problems of consistency and censorship that plague reactive approaches'.[239]

RECOMMENDATIONS TO **STATES**

> Protect freedom of thought as an absolute right in international human rights law.[240]

> **Consider limiting micro-targeting through data privacy laws.**

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> Digital platforms should allow independent audits by vetted researchers on the impact of micro-targeting, in order to understand whether such frameworks should be controlled or limited.

237 Schiffrin, A. (2020). PhD dissertation (University of Navarra).
238 Alegre, S. (2017). Rethinking Freedom of Thought for the 21st Century. *European Human Rights Law Review*
239 Campbell, A. (2019). How Data Privacy Laws Can Fight Fake News. *Just Security.* Retrieved from https://www.justsecurity.org/65795/how-data-privacy-laws-can-fight-fake-news/ (Accessed on October 12, 2020).
240 Alegre, S. *op. cit.*

4.3. STOP FRIENDS-OF-FRIENDS

Digital platforms make recommendations to users in order to connect with other users and groups, based on the digital platforms' assumptions about who users are, what interests they share with other users, and similar criteria.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Stop recommending or otherwise amplifying groups or content from groups associated with hate, misinformation or conspiracies to users.[241]

> Digital platforms should allow independent audits by vetted researchers on the impact of design features aimed at making social interconnections, in order to understand whether such frameworks of social recommendation should be controlled or limited.

4.4. GIVE CHOICE TO USERS

As discussed in Chapter 1, users of digital platforms have a right to know how information is distributed to them, how information targets them, and according to what criteria this occurs.

Currently, digital platforms make assumptions as to what information interests users, and do so in a totally opaque manner. In other words, companies decide who users are without giving them any say in the matter.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> **Give users the choice, in an easy and intelligible manner, to decide on such issues as what information they want to see, and how they are targeted.**
The right to choose who they are and what information they want to see, expands into the right to choose who the platform recommends them to connect with.

241 Stop Hate for Profit. (2020). Recommended Next Steps.
Retrieved from https://www.stophateforprofit.org/productrecommendations (Accessed on October 15, 2020)

Chapter 4:
Mixed Private and Public Spaces on Closed Messaging Services

The virality of disinformation shared on messaging apps is reinforced by the use of groups that sometimes have thousands of members. It is important to define minimal rules for messaging apps that exploit the possibilities of the online public domain while complying with international standards on freedom of opinion and expression.

Contents

INTRODUCTION

Recent developments have shown the rise of closed messaging services as a means of communicating and exchanging ideas, opinions and information. **In 2019, 2.52 billion mobile phone users accessed over-the-top messaging apps to communicate.** In January 2020, 1.6 billion users were active each month on WhatsApp, sending more than 55 billion messages every day, while Facebook Messenger, WeChat and Telegram totaled respectively 1.3 billion, 1.15 billion and 400 million users worldwide.[242] **In some countries like Brazil, India and Spain, closed messaging services have even become the primary source for information.**[243] In Africa, WhatsApp and Facebook Messenger are among the most popular apps.[244]

Over the years, messaging services have gone beyond small groups supporting existing relations between different types of identified individuals—such as family members, friends, colleagues—to large groups that non-members can discover and join, as long there is room in the group.

Although users benefit from such communication systems for their private correspondence, the size of certain groups suggests that the content passing through these services is neither private nor confidential. These new services have blurred the lines between private and public communications and differ from traditional private correspondence. The ability for users to form large groups, coupled with platforms' viral sharing mechanisms, make these services effective tools in spreading misleading information. Such information can reach a large portion of the network and influence public opinion without any sort of moderation or means of decreasing virality. While end-to-end encrypted messaging services, where only intended recipients can read the content of messages, offer strong privacy and security for users and are necessary for the exercise of the right to freedom of opinion and expression in the digital age, they pose challenges to limiting the spread of misleading information. It is important to note that creating vulnerabilities or constraints on encryption is problematic and inconsistent with human rights standards.[245]

Numerous examples confirm the potential for such tools to be abused. In 2018, in India, false rumors about child kidnappers disseminated over WhatsApp fueled mob lynchings that led to more than 20 people being killed over a period of two months.[246] In Brazil, during the 2018 elections, WhatsApp was used to spread false rumors, manipulated photos and decontextualized videos on a large scale.[247] More recently, during the COVID-19 pandemic, closed messaging services were used to spread inaccurate information. WhatsApp was used to spread 5G conspiracy theories, which resulted in people attacking telecoms infrastructure in the United Kingdom.[248] LINE was used by the Malaysia-based operation Qiqi News Network to create and spread COVID-19 disinformation on LINE groups.[249]

---

242 On 25 January 2020, based on monthly active users, active user accounts, advertising audiences, or unique monthly visitors. Slide 95 of Kemp, S. (2020). Digital 2020: Global Digital Overview. Data Reportal. Retrieved from https://datareportal.com/reports/digital-2020-global-digital-overview (Accessed on 1 September 2020).

243 Wardle, C. (2019). Monitoring and Reporting Inside Closed Groups. Verification Handbook For Disinformation And Media Manipulation. (Chapter 7). Data Journalism. Retrieved from https://datajournalism.com/read/handbook/verification-3/investigating-platforms/7-monitoring-and-reporting-inside-closed-groups-and-messaging-apps (Accessed on 21 October 2020).

244 Boyd, C. (2019). WhatsApp in Africa. The Startup. Retrieved from https://medium.com/swlh/whatsapp-in-africa-3c8626f4980e (Accessed on 21 October 2020).

245 Kaye, D. (2015). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.* Retrieved from https://freedex.org/wp-content/blogs.dir/2015/files/2015/10/Dkaye_encryption_annual_report.pdf (Accessed 17 August 2020). para 56.

246 Hern, A. (2020). WhatsApp to Restrict Message Forwarding after India Mob Lynchings. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings (Accessed on 17 September 2020).

247 Boadle, A. (2018). Facebook's WhatsApp Flooded with Fake News in Brazil Election. Reuters. Retrieved from https://www.reuters.com/article/us-brazil-election-whatsapp-explainer-idUSKCN1MU0UP (Accessed on 17 September 2020).

248 Parveen, N. & Waterson, J. (2020). UK Phone Masts Attacked Amid 5G-Coronavirus Conspiracy Theory. *The Guardian.* Retrieved from https://www.theguardian.com/uk-news/2020/apr/04/uk-phone-masts-attacked-amid-5g-coronavirus-conspiracy-theory (Accessed on 10 September 2020).

249 Coca, N. (2020). Disinformation from China Floods Taiwan's Most Popular Messaging App. *Coda Story.* Retrieved from https://www.codastory.com/authoritarian-tech/taiwans-messaging-app/ (Accessed on 13 October 2020).

Recently, **closed messaging services started addressing these issues by putting in place various self-regulatory measures aimed at adding friction in exchanges.** These new frictions, introduced through changes in their features, are designed to disincentivize, disrupt, and slow down the dissemination of information across the network, allowing users to analyze and reflect on the content they share. Introducing frictions into exchanges slows down the dissemination of misleading information throughout the network to the benefit of information reliability.[250]

However, **the arbitrary decisions of these platforms in the way they structure private and public communications call for enhanced democratic supervision.**

1. DEFINING CLOSED MESSAGING SERVICES

A closed messaging service is a service that is provided using an electronic communications network with, as a principal purpose or as an essential functionality of the service, a dissociable section that offers a two-way or multi-way communications component between a finite (that is to say not potentially unlimited) number of natural persons. This number is determined by the sender of the communication.

A closed messaging service is not a service for which the provider has editorial responsibility of, or editorial control over, the content included in the service, and is not a one-to-one telephony service.

Closed messaging services include applications such as Facebook Messenger, Kakao Talk, LINE, Signal, Telegram, WhatsApp and WeChat. They have a strong two-way or multi-way communication component, deliver and display user-generated content to a large member/user audience. This chapter does not cover services which are already subject to a detailed existing regulatory regime.

One-on-one communication—any conversation between two individuals—is not the subject of this report, which focuses on group and viral discussions. One-on-one communication falls under the model of Article 17 of the International Covenant on Civil and Political Rights and Article 8 of the European Convention on Human Rights.

Usually, closed messaging services offer a variety of services including one-on-one conversations, the ability to form groups, channels (characterized by one-to-many conversations), the ability to spread content and forward received messages to a large audience, and the ability to create bots channels that users can subscribe to.

Many closed messaging services are characterized by the use of encryption, which protects communications through a mathematical process that makes the message unreadable except to the receiver and the sender, who have the key to decrypt it into readable form. This is the case, for example, with the Secret Conversation feature of Facebook Messenger, Signal, Telegram and WhatsApp.

Unlike social networks, messaging applications are not driven by self-selection. They do not offer recommendations or an algorithmic timeline that could expose created content to users who are not subscribed to particular channels.

The size of groups differs, depending on the service, but their large scale suggests that the exchange goes beyond private communications and that content exchanged is neither private nor confidential. In

250 Goodman, Ellen (2020). Digital Information Fidelity and Friction. *Knight First Amendment Institute at Columbia University.* Retrieved from https://knightcolumbia.org/content/digital-fidelity-and-friction (Accessed on 8 October 2020). It discusses new sources of friction in information flows to foster information fidelity on digital platforms.

short, closed messaging services simultaneously host private correspondence, content falling within the scope of mass distribution, and even sometimes content distributed to large audiences in the form of private correspondence.

Because of the fluid nature of these spaces, this report does not intend to characterize them, but rather to highlight minimum rules that service providers and states should follow in order to contain the spread of misinformation and disinformation, while complying with international standards on freedom of opinion and expression.

RECOMMENDATIONS
TO **STATES**

> At national levels, regulatory authorities that are independent and whose decisions should be based on the application of internationally recognized principles should monitor and assess whether service providers respect their legal obligations under national regulation. This independent national regulatory authority could be an existing one already in charge of regulating the internet, with a broadened mandate, or a new independent entity created by states.

> Legal obligations, defined throughout this chapter, should apply to closed messaging services that meet an established set of criteria and thresholds. These should be considered further by the Forum on Information & Democracy, and be defined in accordance with civil society.

2. PREVENTING VIOLATIONS TO HUMAN RIGHTS

Similarly to social media platforms, closed messaging services make design choices that may encourage or discourage certain behaviors of the users of their service.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

Prevent violations to human rights

> **Apply the *United Nations Guiding Principles on Business and Human Rights***, and integrate human rights into their products and systems by design and by default, through the development and implementation of their policies, procedures and processes.[251]

251 Kaye, D. (2019). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. OHCHR. Retrieved from https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf (Accessed on 17 August 2020). para 42.

> Include human rights protections in any new policies and services, rather than relying on a model of scaling up first and addressing abuses later.[252]

> **Consult third-party human rights experts and civil society organizations regularly, especially before launching new products, features, or services,**[253] and when reviewing existing policies, to develop relevant technical and organizational measures to be implemented in order to prevent, address and remedy the risks and adverse effects of their activities .

Evaluate impact

> When evaluating the risks, perform, in particular, participatory and periodic public evaluations to determine how content moderation decisions (i.e., decisions concerning the legality of content or its compliance with the terms of service of the service provider, and the operations that result from this decision) are having an impact on the fundamental rights of users, and take the necessary steps to mitigate any harm.

> Share information proactively with researchers and civil society to allow them to independently evaluate the impacts on human rights of content moderation decisions.[254]

> Contribute, including through economic support, to the work of researchers and civil society groups performing independent evaluations.[255]

> Develop and incorporate in a transparent manner, human rights impact-evaluation protocols into their operations to streamline the work of researchers, civil society, and regulators.[256]

252 Pírková, E. & Pallero, J. (2020). 26 Recommendations on Content Governance, a Guide for Lawmakers, Regulators and Company Policy Makers. *Access Now*. Retrieved from https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (Accessed on 22 September 2020). p. 36.
It proposes recommendations to prevent human rights harms on digital platforms. Specific recommendations that could be applied to closed messaging services are listed in this section.
253 Pírková, E. & J. Paller. *op. cit*. p.36
254 Ibid.
255 Ibid.
256 Ibid.

3. RESPECTING THE PRINCIPLES OF FREEDOM OF EXPRESSION WITH REGARD TO THE MODERATION OF ILLEGAL CONTENT

When creating policies and regulations for closed messaging services, states and service providers should ensure any measure taken is necessary and proportionate to the issue being addressed.[257] **Service providers should not have a general and proactive obligation to monitor content proactively.**

When notified, service providers ban users from their service on their own terms, sometimes in violation of the principles of freedom of expression. Service providers should therefore never be allowed to block access to their service without justification, transparency and accountability.

Several organizations have formulated recommendations to states and online service providers to improve the definition of terms of service, reinforce notification mechanisms and appeal mechanisms, and limit the adverse effects of automated moderation techniques on social media platforms. Building on the reports of the UN special rapporteur on the promotion and protection of the right to freedom of expression and opinion, and the work of civil society organizations such as Access Now,[258] Center for Democracy and Technology,[259] EDRi,[260] and Global Partners Digital,[261] similar recommendations and mechanisms are proposed below for closed messaging services.

3.1. DEFINING CLEAR TERMS OF SERVICE

Service providers define terms of service[262] to frame the use of their services, laying out various behaviors and types of content that are not allowed on their platforms.[263]

257 The International Principles on the Application of Human Rights to Communications Surveillance. (2013). Retrieved from https://necessaryandproportionate.org/principles/ (Accessed on 10 September 2020).
258 Pírková, E. & Pallero, J., *op. cit.*
259 Center for Democracy and Technology. (2017). Mixed messages? The Limits of Automated Social Media Content Analysis. Retrieved from https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf (Accessed on 21 August 2020).
260 EDRi contribution to the working group on infodemics.
261 Global Partners Digital contribution to the working group on infodemics.
262 Terms used can differ from one service provider to another: Terms of Service/Terms of Use or Community Standards/Guidelines.
263 See for example:
WhatsApp. (2018). Terms of Service. Retrieved from https://www.whatsapp.com/legal/#terms-of-service (Accessed on 23 September 2020).
Telegram. (n. d.). Terms of Service. Retrieved from https://telegram.org/tos (Accessed on 23 September 2020).
Viber. (n. d.). Conditions d'utilisation de Viber.
Retrieved from https://www.viber.com/terms/viber-terms-use/ (Accessed on 23 September 2020).
Line (n. d.). LINE Terms and Conditions of Use.
Retrieved from https://terms.line.me/line_terms/ (Accessed on 23 September 2020). Kakao Talk (n. d.). Kakao Operation Policy.
Retrieved from https://www.kakao.com/policy/oppolicy?lang=en (Accessed on 23 September 2020).
Kakao Talk (n. d.). Kakao Comprehensive Terms of Service.
Retrieved from https://www.kakao.com/en/terms (Accessed on 23 September 2020).

RECOMMENDATIONS
TO **STATES**

> States should encourage the relevant international institution to state the principle that the service providers' terms of service should be established on the basis of criteria consistent with international standards on freedom of expression, in particular those defined in Article 19 of the International Covenant on Civil and Political Rights, as interpreted by the Human Rights Committee in its General Comment No. 34.[264]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Service providers should develop clear terms of service that are easily accessible and written in clear and understandable terms and available in all languages and dialects where the service is offered (see chapter 1 and Chapter 2).

> Terms of service must comply with international human rights law and standards.

> Terms of service must specify clearly what types of content and activities are prohibited on the provider's services.

> Service providers must update the users when those terms of service are updated.

> Terms of service must clearly specify which types of content moderation operations (banning a user, closing a conversation, etc.) may be exercised, in which case and for what reasons.

3.2. REINFORCING NOTIFICATION MECHANISMS OF ILLEGAL CONTENT

Given the power of closed messaging applications to foster communication between multiple individuals, a user who receives a message that he or she believes contains harmful content must have the ability to flag that message to the service provider. Notification mechanisms already exist on most closed messaging services, however there are discrepancies between them.

Moderation decisions taken by service providers remain opaque and can be executed without warning. For instance, on WhatsApp, any account activity believed to violate the terms of service results in the user being banned from the service.[265] Most of the time, users receive no notifications apart from the following message when trying to use the application: 'Your phone number is banned from using WhatsApp. Contact support for help.'[266]

264 United Nations Human Rights Committee. (2011). General comment No. 34. Article 19: Freedoms of opinion and expression. Retrieved from https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf (Accessed on 10 September 2020).

265 In some cases the blocking can be temporary. Users using unofficial apps developed by third parties to access WhatsApp can be temporarily banned from their service. In this case users receive "an in-app message stating [their] account is 'Temporarily banned'. See, WhatsApp. (n.d.). *About Temporarily Banned Accounts*. Retrieved from https://faq.whatsapp.com/android/temporarily-banned/about-temporarily-banned-accounts/?lang=en (Accessed on 10 September 2020).

266 WhatsApp. (n. d.). Account and Profile. Retrieved from https://faq.whatsapp.com/general/account-and-profile/seeing-the-message-your-phone-number-is-banned-from-using-whatsapp-contact-support-for-help/?lang=en (Accessed on 10 September 2020).

For end-to-end encrypted messaging services, specific recommendations are formulated in part 5.2 of this chapter.

RECOMMENDATIONS TO **STATES**

Notice-and-action procedures

> **At national level, the law should require service providers to put in place adequate notice-and-action procedures. An international entity such as the Forum on Information & Democracy, could be tasked with defining principles and standards for notice-and-action procedures. Details should be specified by the independent national regulatory authorities, in accordance with civil society.**

> Compliance with international standards on freedom of expression implies that the service providers, under penalty of sanction, act against illegal content notified to them, and that they preserve legitimate content.

> Users who have been banned should be able to challenge this removal, possibly with the service provider at first instance, with the possibility of appealing the decision to an independent regulator, under the supervision of a judge.

> Different types of illegal online content and activities may require different responses, specifically tailored to the type of user-generated content that they are supposed to tackle. However, the law has to clearly define the procedures and provide appropriate safeguards for their application by service providers.[267]

> States must regularly evaluate the possible unintended effects of any restrictions before and after applying particular notice-and-action procedures.[268]

Notice-and-review mechanisms

> Conditions for the use of counter-notifications (see below) specified in the law should not be too demanding, because that could discourage users whose content have been moderated from using this mechanism. The law needs to specify what type of content and situation may lead to an exception to the use of counter-notices.[269]

> To counter abusive notifications and reports in bad faith, the law should require service providers to respect a set of transparency obligations, which are defined in part 6 of this chapter.

> A regime against abusive and bad faith reporting should be established. Procedures to guard against misuse of reporting rules and moderation mechanisms should be required from service providers.

267 Pírková, E. & J. Pallero. *op. cit.* p.27
It proposes detailed recommendations for notice-and-action procedures on online platforms in the case of illegal content notified to them.
268 Ibid., p.28
269 Ibid., p.32
It defined counter-notifications as a mechanism that "enables content providers to object to individual complaints targeting their content."

RECOMMENDATIONS TO **SERVICE PROVIDERS**

Setting up notification mechanisms

> Service providers must ensure they have the functionality to allow users to easily notify them of content which they consider to be in breach of the terms of service and applicable laws.

> These mechanisms must be visible, easily accessible and user-friendly. They should not be unnecessarily time-consuming to discourage users from using them.

> To make notification mechanisms more effective, service providers should provide a detailed form for the notification with, in particular, a list of reasons for submitting a notice, and request information on the context of the notified content.

> People who are not members of groups that can be joined via a URL should be allowed to report groups to the service provider (when applicable) if the group's name has a negative indication. Groups whose titles violate international standards of freedom of expression (e.g. contains calls for hatred, violence, etc.) could be shut down.

Create counter-notifications mechanisms

> Any notification of content must be subject to information to the author or members of the group, allowing them to contest this notification ('counter-notification').

Create notice-and-review mechanisms

> Service providers should notify both the user who flagged the content and the user who uploaded or generated the content when a moderation decision has been made about their content or speech, and the notification should contain the requisite information to ask for a review of the decision.

> The notification should contain, at least:
- the reasons for temporarily or permanently banning users, including the specific rule that has been infringed and how the content moderation guidelines were interpreted;
- a precise explanation of the content provider's rights;
- a clear explanation of possibilities for appealing the decision;
- the clearly stated option of judicial redress.

> The sanctions applied by service providers should be proportionate to the offence committed by the user. International human rights law demands that limitation of expression should only be that which is necessary to achieve a legitimate purpose.[270] Restrictions should not be overbroad; the least intrusive means should be used to limit expression[271] (see part 1.2 of Chapter 2).

270 Article 19(3), ICCPR.
271 United Nations Human Rights Committee, General Comment no. 34: Article 19: Freedoms of Opinion and Expression, CCPR/C/GC/34 (12 September 2011), para. 34.

3.3. REINFORCING APPEAL MECHANISMS

Users of closed messaging applications have little possibility of contesting service providers' decisions on content moderation.

RECOMMENDATIONS
TO **STATES**

> **States should impose the establishment of internal appeal mechanisms on service providers.**

> **Service providers should be required to inform the authors of banned accounts of the possibility of contesting these decisions, and of the possibility of challenging a service provider's decision on this appeal in court.**

> Once the internal mechanisms of closed messaging services are exhausted, service providers should have an obligation to inform the user about the legal remedies available to them in their national jurisdiction. Users should have the ability to appeal this decision before an independent regulator or in court.

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Users should have a possibility of effectively appealing against a service provider's decision. The recourse process should be easily understandable and accessible.

3.4. AUTOMATED MODERATION TECHNIQUES

Legislative initiatives aimed at regulating online service providers often foster the use of automated and AI-driven analysis of content. Similarly, closed messaging services are subject to the same pressure from regulatory authorities.

Automated moderation techniques on platforms are used to counter the spread of harmful content, especially of terrorist and violent extremist content and child sexual abuse material. However, these techniques are limited in a variety of ways, and can be unreliable for assessing content which needs more context and nuance of language. Moreover, the technology in itself can present structural biases, resulting in errors and over-deletion of content.

RECOMMENDATIONS
TO **STATES**

> **The use of automated tools should not result in a general monitoring obligation of communications by service providers.**

> Use of automated content analysis tools to detect or remove illegal content should never be mandated in law.[272]

> States should be aware of the technical limitations of automated moderation technologies and should understand the vulnerabilities arising from reliance on artificial intelligence and assumptions embedded in its design-based user interfaces.[273]

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Be transparent about the use of automated tools for content moderation operations. These tools should operate and use criteria that are in line with international human rights standards.

> To avoid false positives, any use of automated content analysis tools should be accompanied by a human review of the output or conclusions produced by the tool.

> The use of automated measures should be accepted only in limited cases of manifestly illegal content that is not context-dependent.[274]

> Develop mechanisms for increased accountability of automated content analysis tools used, including civil society oversight for any information-sharing models that implicate digital rights (see part 6 of this chapter for more details).[275]

4. LIMITING VIRALITY OF CONTENT

Even though closed messaging services do not offer recommendations or curation systems that push content forward to users, some form of virality exists on these apps. **Closed messaging services should continue adopting measures that limit the virality of false or misleading content shared on their networks without undermining the rights to privacy and freedom of opinion and expression of their users.**

4.1. ADDING FRICTION

Adding friction to limit the spread of misleading information throughout a service network appears to be a solution that has been taken lately by several companies. These new frictions, introduced through changes in features offered by messaging apps, are designed to disincentivize, disrupt, and slow

272 Center for Democracy and Technology. (2017). *op. cit.* p.6.
273 Heller, B. (2019). Combating Terrorist-Related Content Through AI and Information Sharing. Transatlantic Working Group. Retrieved from https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf (Accessed on 29 September 2020). p.7.
274 Pírková, E. & J. Pallero. *op. cit.* p.31.
275 Heller, B., *op. cit.* p.6.

down the flow of information across the network, allowing users to analyze and reflect on the content they share. Introducing frictions into exchanges slows down the dissemination of misleading information throughout the network to the benefit of information reliability.[276]

Service providers have already implemented several measures:

They limit the number of users in groups to prevent the creation of giant hubs. Service providers have imposed varying limits: iMessage – 20; Facebook Messenger – 150; WhatsApp – 256; LINE – 500; Telegram – 200,000. However these limitations do not prevent content from reaching a large proportion of the platform.[277]

They limit the ability for users to forward messages to more than a certain number of groups at a time to slow down the spread of messages. Early in the COVID-19 pandemic, WhatsApp imposed new limitations on the ability to forward messages. Users who received a frequently forwarded message—one which had been forwarded more than five times—were able to send it on only to a single chat at a time.[278] In September 2020, Facebook Messenger also reduced the number of forwards to five groups at a time.[279] A study, based on public data, has shown that imposing low limits on forwarding in comparison with the original limit of 256 used in the first version of WhatsApp, offers a delay in the message propagation of up to two orders of magnitude. When the forwarding limit was reduced to five, 80 percent of messages died within two days. However, 20 percent were still very viral and reached the full network during this time.[280] Imposing low limits on forwarding offers a delay in message propagation, but depending on the virality of the content such limits are not effective in preventing a message reaching the entire network quickly.[281] It also depends on the service provider. Another study looking into the propagation of 'junk news' in English on Telegram has found that sharing a message between channels does not significantly increase the audience.[282]

They also combat automated behaviors,[283] provide settings so users can configure who can add them into groups,[284] and implement opt-in features for users when they are added in groups.

However, because these closed messaging services combine viral sharing mechanics with private correspondence functionalities sometimes using end-to-end encryption, additional measures are proposed below.

276 Goodman, Ellen (2020). *op. cit.*
It discusses new sources of friction in information flows to foster information fidelity on digital platforms.

277 De Freitas Melo, P., C. Coimbra Vieira, K. Garimella, P. O. S.Vaz de Melo, & F. Benevenuto. (2019). Can WhatsApp Counter Misinformation by Limiting Message Forwarding.
Retrieved from https://arxiv.org/pdf/1909.08740.pdf (Accessed on 23 September 2020).

278 Hern, A. (2020). WhatsApp to Impose New Limit on Forwarding to Fight Fake News. *The Guardian*.
Retrieved from https://www.theguardian.com/technology/2020/apr/07/whatsapp-to-impose-new-limit-on-forwarding-to-fight-fake-news (Accessed on 7 October 2020).

279 Sullivan, J. (2020). Introducing a Forwarding Limit on Messenger. Facebook Newsroom.
Retrieved from https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/ (Accessed on 7 October 2020)

280 Chen, A. (2019). Limiting Message Forwarding on WhatsApp Helped Slow Disinformation. *MIT Technology Review*.
Retrieved from https://www.technologyreview.com/2019/09/26/434/whatsapp-disinformation-message-forwarding-politics-technology-brazil-india-election/ (Accessed on 29 September 2020).

281 De Freitas Melo, P., et al., *op. cit.*

282 Knuutila, A.., A. Herasimenka, J. Bright, R. Nielsen, & P. N. Howard. (2020). Junk News Distribution on Telegram.
Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/07/Junk-News-Distribution-on-Telegram.-Data-Memo.pdf (Accessed on 23 September 2020).

283 See for instance WhatsApp measures to combat abuses of its services.
WhatsApp (2019). Stopping Abuse: How WhatsApp Fights Bulk Messaging and Automated Behavior. Retrieved from https://scontent.whatsapp.net/v/t61.22868-34/69510151_652112781951150_6923638360331596993_n.pdf/Stopping-Abuse-white-paper.pdf?_nc_sid=2fbf2a&_nc_ohc=8kN5Smy85pYAX-yQdzz&_nc_ht=scontent.whatsapp.net&oh=80a9f74996b72496ebbb84fa27787c2b&oe=5F872CCF (Accessed on 13 October).

284 *The Economic Times*. (2019). Whatsapp Users Can Now Decide Who Can Add Them To Groups.
Retrieved from https://economictimes.indiatimes.com/tech/internet/whatsapp-users-can-now-decide-who-can-add-them-to-groups/articleshow/71947671.cms?from=mdr (Accessed on 13 October 2020).

RECOMMENDATIONS TO **STATES**

In order to preserve usefulness of the service as a closed messaging service while making its exploitation more difficult in order to limit the potential for abuse of its features, some functionalities should be limited:

> States should impose a legal obligation on service providers to set limits to service providers' features (for example set limits on the maximum number of users allowed in groups and limits to forwarding features). An international entity, such as the Forum on Information & Democracy, could be tasked with defining appropriate limits. The actual specifications could be determined by independent national regulatory authorities as defined in part 1 of this chapter.

> The independent national regulatory authorities would monitor that companies respect their obligations based on information provided by service providers (see part 6 of this chapter).

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Comply with legal obligations regarding limitations to their service features (see above).**

> **Implement other relevant measures such as:**

- **Require users to 'opt in' to receiving group messages**[285] and before their personal information (names, profile pictures, phone numbers) are revealed to the rest of the group.
- **Allow users to choose who can add them into groups** through settings allowing users to define what kinds of contact can add them. By default, this option should be set so that only contacts from the user's phone book can add the user into groups.
- **Limit the forwarding feature** to reduce the risk of abuse of this functionality. Service providers could restrict users to forwarding to one chat group at a time in order to preserve platforms' usefulness as a closed messaging service while making it more difficult to exploit.[286]
- **Label messages created by bots or messages sent by business accounts.**[287]
- **Fight bulk messaging and automated behavior** by banning the use of external tools not approved by the service provider to manage accounts.

285 Avaaz contribution to the working group on infodemics
286 Barrett, P. M. (2019). *Disinformation and the 2020 Election: How the Social Media Industry Should Prepare.* NYU Stern Center. Retrieved from https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_election_2020_report?fr=sY2QzYzI0MjMwMA (Accessed on 29 September 2020).
287 Avaaz contribution to the working group on infodemics

4.2. EMPOWERING THE USER

Providing users with credible sources of information and ways to make informed decisions is crucial. Service providers have already implemented some measures in this respect.

They label messages as 'forwarded' to indicate they did not originate from a close contact. This feature, launched by WhatsApp in mid-2018, adds a single arrow icon along with the word 'Forwarded' to every forwarded message,[288] and from mid-2019, a double-arrow icon to mark messages that are at least five forwards away from the original sender.[289] However, the design of the application does not provide detailed information on what this icon means.

They create mechanisms to facilitate the verification of information forwarded to users. WhatsApp recently launched a feature giving users the ability to easily export a message to their browser to search for more context.[290] However, usage of the Internet differs in different regions of the world. Due to the high cost of data, some users purchase specific Internet bundles valid for a given duration of time. These are widely used to access WhatsApp. Such bundles cover anything received on Facebook and WhatsApp, but do not allow users to go on the Internet to carry out additional research, or to access the content of websites shared in conversations.

They allow the creation of bots that can be used to provide users with information. Many closed messaging services provide this feature, such as WhatsApp, LINE, Telegram and Viber. Several governments—such as that of India,[291] which has launched an automated chatbot on WhatsApp, and Serbia[292] on Viber—have created bots to share necessary information about the pandemic, create awareness about COVID-19 and answer questions asked by users about the pandemic. Civil society organizations and fact-checking agencies have launched similar bots, such as CoFacts[293] on LINE or Comprova[294] on WhatsApp. Bots are easily accessible for users, who can send a message to a specific number or click on a link to start receiving updates and ask specific questions. They are also easy for organizations to set up. These channels can be particularly useful, especially in the context of the COVID-19 pandemic. A study[295] conducted in Zimbabwe in the context of the COVID-19 pandemic found that WhatsApp messaging from a trusted source aimed to counteract misinformation can increase both knowledge about COVID-19 and also preventative behavior. This confirms the important role that trusted sources play in combating misinformation in confusing international situations.

However, bots alone do not represent a sufficient solution to promote the reliability of information on closed messaging services.

288 WhatsApp Blog (2018). Labeling Forwarded Messages. Retrieved from https://blog.whatsapp.com/labeling-forwarded-messages (Accessed on 24 August 2020).
289 WhatsApp FAQ. About forwarding limits. Retrieved from https://faq.whatsapp.com/general/coronavirus-product-changes/about-forwarding-limits (Accessed 30 September 2020).
290 WhatsApp Blog. (2020). Search the Web. Retrieved from https://blog.whatsapp.com/search-the-web (Accessed on 23 August 2020).
291 Singh, M. (2020). India Launches Whatsapp Chatbot to Create Awareness About Coronavirus, Asks Social Media Services to Curb Spread of Misinformation. *TechCrunch*. Retrieved from https://techcrunch.com/2020/03/21/india-whatsapp-mygov-corona-helpdesk-bot/ (Accessed on 23 August 2020).
292 Viber. (2020). COVID-19 Info Srbija. Retrieved from https://chats.viber.com/covid19info (Accessed on 13 October 2020).
293 Coca, N. (2020). *op. cit.*
294 First Draft News. (n. d.). Comprova. Retrieved from https://firstdraftnews.org/project/comprova/ (Accessed on 13 October 2020).
295 Bowles J., H. Larreguy, & S. Liu. (2020). *Countering Misinformation via WhatsApp: Evidence From the COVID-19 Pandemic in Zimbabwe.* Center for International Development at Harvard University. Retrieved from https://www.hks.harvard.edu/sites/default/files/centers/cid/files/publications/faculty-working-papers/2020-05-CID-WP-380-Countering%20Misinformatoin%20Via%20WhatsApp-Evidence%20from%20the%20COVID-19%20Pandemic%20in%20Zimbabwe.pdf (Accessed on 3 September 2020).

RECOMMENDATIONS TO **STATES**

> States should impose on service providers:

- **An obligation to inform users on any new features that have an impact on the design of an app, in all languages and dialects where the service operates.** This could be done through notifications, or a specific channel in the app where the service provider can send updates to users.
- **An obligation to protect users' freedom of opinion by providing them with ways to make informed decisions on their services.** States should consider to this end measures such as obligations to:
  - **Labeling messages as 'forwarded' to indicate they did not originate from close contact.** When labeling messages as 'forwarded', service providers should provide detailed explanations of the imagery used.
  - **Labeling messages created by corporate accounts or bots.**
  - **Creating mechanisms to facilitate the verification of information forwarded to users**. Allowing users to access more contextual information is essential, but these mechanisms should not reveal the content of the message to the service provider.
- **When imposing obligations on platforms to protect users' freedom of opinion, States must always ensure such measures cannot negatively impact the rights to freedom of opinion and expression.**

> The national regulatory authorities should monitor that companies respect their obligations based on information provided by service providers (see part 6 of this chapter).

RECOMMENDATIONS TO **SERVICE PROVIDERS**

Implement other relevant measures such as:

> **Provide features that enable the user to define the character of the message and limit its ability to become viral.** If users can define whether their message can be forwarded, they can state whether or not they want to allow the message to go viral. This defines a clear threshold between interpersonal and viral messages, and allows the application of proportional measures to track authors in the event of illicit viral content.

> **Restrict the use of bots whose purpose is to respond to requests for information** to certain types of topics such as scientific information, health information or voting rules. The bots should not become a means, for example, to check political speech.

> **Create specific programs to subsidize the establishment of bots** by fact-checking agencies, local civil society organizations and the media. The sources of information promoted should meet ethical and professional standards for the production of reliable information. Their identification should be based on a standard such as the Journalism Trust Initiative, established under the aegis of European Committee for Standardization (CEN) and in cooperation with journalists and media from all over the world (see section 2.2. of Chapter 3 for more details).

> **Allow users to report photos received in conversations to a specific channel of the service.** Service providers could then analyze the picture against a database of pictures debunked by independent fact-checking agencies and send the user all available information about the picture (such as the date it was taken or published/ distributed, the location, etc.).

5. PROTECTING END-TO-END ENCRYPTED MESSAGING SERVICES

Recently, some states have proposed worrying legislation aimed at regulating closed messaging services offering end-to-end encrypted communications. Recent cases include the Brazilian 'Fake News Law',[296] which include an obligation for service providers to track all chains of communication and retain the metadata; the proposed amendments to the Indian Information Technology (Intermediaries Guidelines) Rules under the Information Technology Act, which would require companies to be able to trace the originator of a given piece of content on their platform,[297] and the EARN IT Act currently discussed in the United States, which would require service providers to identify and take down child sexual abuse material, an obligation that could result in the implementation of client-side scanning of content in order to check the fingerprint (also called hash) of images and videos sent by users against a dataset of known harmful content.[298]

Access to message content in the context of crime solving and terrorism prevention is an issue that will not be addressed in this chapter. This issue could be the subject of a future working group of the Forum on Information and Democracy.

296 On 30 June 2020, Brazil's Senate passed the Draft Bill no. 2.630, of 2020 (PL 2630/2020), known as the 'Fake News Law', to combat the spread of disinformation on online social networks, and which includes obligations to trace communications. See
Aleixo, G., A. Guimarães Gobbato, I. Garcia de Souza, N. Langenegger, R. Lemos, & F. Steibel. (2019). The Encryption Debate in Brazil. Carnegie Endowment for International Peace. Retrieved from https://carnegieendowment.org/2019/05/30/encryption-debate-in-brazil-pub-79219 (Accessed on 20 September 2020).
Hartmann, I. A., Y. Curzi, J. Iunes, L. Abbas, & B. Diniz. (2020). Draft Bill no. 2.630, of 2020. Retrieved from https://docs.google.com/document/d/1MHMDHsVJBi45PI1R5lAyoLmZvZk8eULHisYFqGy9X2s/edit (Accessed on 20 September 2020).

297 In December 2018, the Ministry of Electronics and Information Technology released the Intermediary Liability Guidelines (Amendment) Rules. Its objective is to fight the spread of disinformation online. Among other concerning obligations, it includes tracing requirements for companies. See
Arun, C. & N. Nayak. (2016). *Preliminary Findings on Online Hate Speech and the Law in India*. Berkman Klein Center, Harvard. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882238 (Accessed on 13 September 2020).
Quay-de la Vallee, H. (2020). *Proposed Indian Internet Regulations Would Harm Global Internet Security.* Center for Democracy & Technology. Retrieved from https://cdt.org/insights/proposed-indian-internet-regulations-would-harm-global-internet-security/ (Accessed on 28 September 2020).

298 "In order for providers to identify and take down child sexual abuse material on their services, the bill requires either a third party to report such content, or the ability of the provider to view each piece of content passing through their service." from
Quay-de la Vallee, H. & M. Azarmi. (2020). *The New EARN IT Act Still Threatens Encryption and Child Exploitation Prosecutions*. Center for Democracy & Technology. Retrieved from https://cdt.org/insights/the-new-earn-it-act-still-threatens-encryption-and-child-exploitation-prosecutions/ (Accessed on 18 September 2020).
Mullin, J. (2020). The New EARN IT Bill Still Threatens Encryption and Free Speech. Electronic Frontier Foundation. Retrieved from https://www.eff.org/fr/deeplinks/2020/07/new-earn-it-bill-still-threatens-encryption-and-free-speech (Accessed on 18 September 2020).

5.1. TECHNICAL CHALLENGES

Such legal obligations imposed by the regulator would require changes in the technical architecture of private messaging applications. **Preserving end-to-end encryption is crucial, as it is a tool essential to the protection of human rights.**[299] Worldwide, journalists and their sources rely heavily on the security of established end-to-end encryption in closed messaging services. While criminals can easily decide to use a different messaging service, establishing an alternative secure communication channel between a journalist and a source is complicated, error-prone, and overall very difficult to achieve.

RECOMMENDATIONS
TO **STATES**

> **Legal obligations imposed by states must not result in the implementation of technical protocols that could weaken end-to-end encryption, create new vulnerabilities, or undermine international standards of freedom of expression and opinion of users.**

> **Any measure imposed by law must be necessary, proportionate and legitimate in relation to the objective pursued.**

5.2. NOTIFICATION MECHANISMS

In an end-to-end encrypted messaging system, a service provider cannot read the content of messages but intended recipients can. If a user receives a message that he or she believes contains content that goes against the terms of service of the company, the user can flag that message to the service provider through a notification mechanism. One of the technical challenges of enabling user reports lies in authenticating the notification to ensure that a specific sender actually sent the flagged message at a specific time.[300] From a scientific point of view, the understanding of the security of these mechanisms is in its infancy.[301]

WhatsApp and the Secret Conversation feature of Facebook Messenger, allowing users to create end-to-end encrypted conversation, both allow users to report conversations. On WhatsApp, a picture of the last messages of the conversation is then taken on the client side and sent to the service provider for evaluation. On Secret Conversation, recent messages from the conversations are 'decrypted and sent securely from [the user's] device to [Facebook's] Help Team for review'.[302] They use a process called the 'franking mechanism'[303] to authenticate the messages received. This allows them to associate each message with a cryptographic statement of authenticity, confirming that it was sent by a particular sender at a particular time, without revealing the content of the message. If a user reports a message to the service provider, the provider can then check the statement of authenticity and verify the flagged message. Facebook will never have access to plaintext messages unless one participant in a secret

299 Global Partners Digital. (2017). Encryption Policy for Human Rights Defenders. Retrieved from https://www.gp-digital.org/wp-content/uploads/2017/09/TRAVELGUIDETOENCRYPTIONPOLICY.pdf (Accessed on 21 August 2020).
300 Mayer, J. (2019). Content Moderation for End-to-End Encrypted Messaging. Retrieved from https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf (Accessed on 29 August 2020).
301 Earliest work found on this topic seems to be
Grubbs, et al. (2017). Message Franking via Committing Authenticated Encryption. Retrieved from https://eprint.iacr.org/2017/664.pdf (Accessed on 23 October 2020).
302 Facebook. Help Center. (n. d.). How do I report a secret conversation in Messenger?
Retrieved from https://www.facebook.com/help/messenger-app/android/498828660322839 (Accessed on 29 August 2020).
303 Facebook. (2017). Messenger Secret Conversations, Technical Whitepapers. Version 2.0. Retrieved from https://fbnewsroomus.files.wordpress.com/2016/07/messenger-secret-conversations-technical-whitepaper.pdf (Accessed on 29 August 2020).

conversation voluntarily reports the conversation. However, research has shown that this mechanism is insecure.[304]

RECOMMENDATIONS
TO **STATES**

> Specify that technical protocols implemented to authenticate reports in service providers' notification mechanisms must comply with a set of safeguards to be considered further and defined in accordance with civil society and independent researchers (see part 7 of this chapter on empowering research).

> States should refrain from engaging in online disinformation and consider measures that encourage companies to fight against disinformation (see below).

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Service providers that offer notification mechanisms should make the protocols implemented to authenticate reports auditable by a regulator, independent auditors and the research community (see part 6 on transparency requirements for more details).

> Service providers that offer notification mechanisms should allow users to report possible disinformation content, even on private and encrypted channels. This would enable service providers to respond to such reports by providing users with reliable sources of information that have been published on the subject that has been reported. The sources of information promoted should meet ethical and professional standards for the production of reliable information. Their identification should be based on a standard such as the Journalism Trust Initiative established under the aegis of CEN and in cooperation with journalists and media from all over the world (see section 2.2. of Chapter 3 for more details).

5.3. KNOWN CONTENT DETECTION SYSTEMS

This technique, also called hash-matching, is already common practice among email services, hosting and social networks and some closed messaging services, such as Facebook Messenger which scans every image that people send to detect known child sexual abuse material.[305] However, on end-to-end encrypted messaging applications, service providers cannot use the same content detection systems because the service provider cannot and should not access user content.

The technical implementation of such mechanisms to identify illegal content such as child sexual abuse material and terrorist and violent extremist content would require service providers to modify the architecture of their applications. Such a change would have several consequences. Among them, the fact that such a system would introduce new vulnerabilities in the service architecture. In addition, the

304 Research published has shown ways to break Facebook's franking mechanism. See Langkemper, S. (2019). Breaking Message Franking. Retrieved from https://www.sjoerdlangkemper.nl/2019/11/20/message-franking/ (Accessed on 23 October 2020). Grubbs, et al. (2019). Fast Message Franking: From Invisible Salamanders to Encryptment. Retrieved from https://eprint.iacr.org/2019/016.pdf. (Accessed on 23 October 2020).

305 Frier, S. (2018). Facebook Scans the Photos and Links You Send on Messenger. Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app (Accessed on 23 August 2020).

verification system, supposed to run on the users' device ('client-side'), would probably run on the server because of the size of the databases, which even in a compressed format would take up too much space.[306] Moreover, such a system would break the promise of end-to-end encryption—the fact that only intended recipients can read and understand the content of a message—since the server would be able to decrypt part of the message.[307] Finally, content detection systems always come with a false-positive and false-negative rate.[308] On unencrypted services, providers can manually investigate any reports and decide if it is a false-positive or indeed a true-positive. For end-to-end encrypted messages, this would require the provider to be able to decrypt the message—even in cases where the testing of images is done solely on the user's device.

Recently, researchers have explored the feasibility of a known content detection system for disinformation based on content debunked by independent fact-checking agencies. Such a system would verify the hash (also called fingerprint) of images and videos sent in a conversation by a user against a database of known debunked images and videos.[309]

RECOMMENDATIONS
TO **STATES**

> **Do not impose known content detection systems that would undermine end-to-end encryption.**

RECOMMENDATIONS
TO **SERVICE PROVIDERS**

> Refuse to implement known content detection systems that would undermine end-to-end encryption. Accept detection of known content only on the basis of a notification.

5.4. THE CONTENT TRACING DEBATE

To circumvent malicious coordinated action on closed messaging services, several regulatory initiatives have emerged to trace the origin of messages sent.[310] These proposals present many flaws. From a technical perspective, such obligations encourage service providers to store massive amounts of metadata about all users' communications, hence to go against the protection of users' right to privacy, as metadata is personal data. Moreover, due to uncertainty over what technical protocols service providers would use to match the traceability obligations, origin-tracing technology could add vulnerabilities to their services, which would weaken the overall security of their applications. Finally, such technical implementation could also undermine end-to-end encryption in itself.[311]

306 Quay-de la Vallee, H., Azarmi, M. (2020)., *op. cit.*
307 Portnoy, E. (2019). Why Adding Client-Side Scanning Breaks End-To-End Encryption. Electronic Frontier Foundation. Retrieved from https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption (Accessed on 18 September 2020).
Crocker, A. & G. Gebhart. (2019). Don't Let Encrypted Messaging Become a Hollow Promise. Electronic Frontier Foundation. Retrieved from https://www.eff.org/deeplinks/2019/07/dont-let-encrypted-messaging-become-hollow-promise (Accessed on 18 September 2020).
308 The most prominent detection system for child sexual abuse material is Microsoft's PhotoDNA. This system is not documented publicly but according to reports, the false-positive rate is between one in one billion and one in ten billion.
309 Reis, J. C. S., P. Melo, K. Garimella, F. Benevenuto. (2020). Can WhatsApp Benefit from Debunked Fact-Checked Stories to Reduce Misinformation?. *The Harvard Kennedy School (HKS) Misinformation Review*. Retrieved from https://doi.org/10.37016/mr-2020-035 (Accessed on 23 August 2020).
310 See part 5.1 of this chapter for more details on Brazil and India regulatory proposals.
311 Rodriguez, K. & S. Schoen. (2020). FAQ: Why Brazil's Plan to Mandate Traceability in Private Messaging Apps Will Break User's Expectation of Privacy and Security. Retrieved from https://www.eff.org/fr/deeplinks/2020/08/faq-why-brazils-plan-mandate-traceability-private-messaging-apps-will-break-users (Accessed on 18 September 2020).
Maheshwari, N. (2020). Traceability Under Brazil's Proposed Fake News Law Would Undermine Users' Privacy and Freedom of Expression. Center for Democracy and Technology. Retrieved from https://cdt.org/insights/traceability-under-brazils-proposed-fake-news-law-would-undermine-users-privacy-and-freedom-of-expression/ (Accessed on 18 September 2020).

RECOMMENDATIONS TO **STATES**

> **Do not encourage the collection, storage and usage of the metadata of all users' communications. Any processing of metadata should be relevant and proportional in relation to the purposes for which the data is processed.**[312]

> Do not criminalize the sharing of information on closed messaging services. Proving intent is complicated, and no one should be punished by law for unintentionally spreading false, inaccurate or misleading information. In addition, the effectiveness of traceability protocols has not yet been proven. This could result in users who are not the original sender being charged.

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> Limit collection, storage and usage of users metadata to legitimate, specific and explicit purposes.[313]

312 Rodriguez, K. & S. Schoen. (2020). *op. cit.*
313 Ibid.

6. CLARIFYING TRANSPARENCY OBLIGATIONS

Similarly to social media, closed messaging services must be subject to clear transparency requirements. The broader requirements are detailed in the first chapter of this report. Specific recommendations are detailed below.

RECOMMENDATIONS TO **STATES**

> **Impose transparency obligations for closed messaging services by law** (see part 1.1 of Chapter 1). The transparency requirements should include **how the service providers implement and follow their own terms of service and internal policies, how technical protocols & algorithms operate and with what objectives, how users' metadata is used, what activities were conducted to identify risks of harm, and the negative effects of their activities on users and other people.**

> The governance of transparency should follow a similar model as the one laid out in part 2 of Chapter 1. The transparency framework should be supervised by an independent regulatory authority. Part 2 of Chapter 1 provides a detailed explanation of how, in different parts of the world, states can implement this accountability regime.

> In order to allow the audit of transparency obligations by independent auditors and researchers and independent regulatory authorities while preserving privacy of users, a similar three-tier system of disclosure could be proposed (see part 2.3 of Chapter 1).[314] Information and data that should be disclosed by service providers are detailed below.

> The same sanctions for non-compliance with transparency requirements should be applied as those defined in part 2.5 of Chapter 1.

Transparency obligations of states

> **States should publish detailed transparency reports on all content-related requests issued to service providers.**[315] These reports should be made publicly available and should include:

- The number and nature of restriction requests issued to service providers. States should justify these requests by providing the legal ground on which the requests were based, including those based on international mutual legal assistance treaties.
- Responses and action taken by service providers following the states' requests.
- Disclosure of all agreements made with service providers.

314 In the three-tier disclosure defined in part 2.3. of Chapter 1, three levels of information are defined: (1) for users; (2) for researchers, independent auditors and regulators; (3) access to restricted data under limited circumstances, such as an investigation.

315 See principle 6.c of the Manilla Principles, https://www.manilaprinciples.org/
Kaye,D. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. United Nations. Retrieved from https://www.undocs.org/A/HRC/38/35 (Accessed on 17 September 2020). para 69.
Pírková, E. & J. Pallero. (2020). *op. cit.*
It proposes a detailed list of types of data that should be included in States' periodic transparency reports.

RECOMMENDATIONS TO **SERVICE PROVIDERS**

Information and data that should be disclosed by service providers at each level of the transparency framework is as follows:

	(1) Public users	**(2) Vetted researchers & regulators**	**(3) Limited access to regulators & researchers approved by regulators**
Nature of information			
Terms of service, community guidelines, internal policies, enforcement guidelines, notification of changes	Yes	Standards of implementation	
Notice of violations of terms of service (TOS) or laws	Users who file a complaint and users whose account was affected	Yes	
Blocked and banned accounts, group takedowns, accounts and groups remaining		Yes	
Notification of users & redress mechanism	If a user's account is banned, or if groups are taken down; with an explanation of redress procedures	Yes	
Technical protocols & automated content moderation techniques		Yes	Yes
Information on use of users' metadata	Yes	Yes	
Human Rights Assessments	Summary	Yes	

Service providers should:

> Make publicly available, in all languages and dialects of the countries where their services are provided, **terms of service, community guidelines, internal policies, enforcement guidelines and notification of changes** (see part 1.2 of Chapter 1 for more details).

> Explain clearly the process for reporting content or groups as in **violation of the terms of service** or laws (see part 3.2 of this chapter and part 1.3 of Chapter 1 for more details).

- Service providers should be transparent about the sanctions applied when users or groups are in violation of their terms of service or laws. They should specify:
  - in which cases a group is shut down;
  - in which cases a user is temporarily or permanently banned;
  - in which cases both of these occurred;
  - the reasons for these sanctions.

> Disclose granular and standardized **data on content-related requests and procedures,** especially regarding blocked and banned accounts, group takedowns, as well as accounts and groups remaining (see part 1.3.a and 1.3.b of Chapter 1). Data should include:

- the number of all notices received, including:
  - the number of one-on-one conversations flagged;
  - the number of groups flagged (with the average number of users in each group);
- the number of accounts and groups that were removed, and a breakdown of reasons why they infringed terms of service:
  - per automated decision, specifying the criteria applied, and releasing the accuracy rate;
  - per human decision, specifying the criteria applied, and releasing the accuracy rate. Service providers should share statistics on human reviewers' inter-rater reliability;
- type of entities that issued the notice, including private parties, administrative bodies, or courts;
- reasons for determining the legality of content, or how it infringes terms of service;
- the number of appeals received and how they were resolved;
- in the case of encrypted messaging services:
  - the number of requests that could not be resolved because of a lack of information.

> Disclose information related to their policies on collecting, storing, retaining, using, selling and sharing **users' data,** including metadata.

- Information should include:
  - what metadata they collect, and for what purpose;
  - where and how the data is stored;
  - how long it is retained;
  - how long it is used for (in content moderation operations, or between the platform's services);
  - with whom data is shared (including between their own services) and under what conditions;
- Users should have the ability to export a report of their account information and

settings.

> **Notify users and give information about redress mechanisms** (see part 3.2 of this chapter for more details).

> **Provide technical protocols & automated content moderation techniques when applicable** (see part 1.4 of Chapter 1 for more details)

- Service providers should clearly and transparently publish meaningful and easily understandable information on what processes are being used, for what purposes, and how decisions are made by those processes.[316]
- The technical protocols and automation, and their results, should be regularly reviewed, and the processes refined, to mitigate against the risks identified above.[317]
- Access can be reserved to auditors who are subject to a confidentiality requirement.

> **Allow transparency on their activities to identify risks of harms and the negative effects of their activities** (See part 1.8 of chapter 1 for more details).

316 Global Partners digital. (2018). A rights-respecting model of online content regulation by platforms. Retrieved from https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf (Accessed on 21 August 2020). p.22.

317 Ibid.

7. EMPOWERING RESEARCH

Closed messaging services remain a rather unexplored field of study. Existing research focuses mainly on WhatsApp, and to a lesser extent on Telegram. This is in part due to the lack of information and data available to researchers, because of the closed nature of these messaging systems. Existing research focuses on data from public groups, meaning URLs to the groups that can be found on the internet,[318] or surveys, focus groups and interviews conducted with local users.[319]

RECOMMENDATIONS TO **STATES**

> Increase funding and support for independent research in the field of closed messaging services to enable evidence-based analysis and policy making.

> **Fund an international research institute focusing on the field of closed messaging services, in order to develop policy recommendations, and conduct further research on options for limiting the virality of disinformation in a rights-respecting manner.**

> Promote independent research and multi-stakeholder collaborations.

> Encourage service providers to conduct research on limiting the virality of disinformation on their services in a rights-respecting manner.[320]

RECOMMENDATIONS TO **SERVICE PROVIDERS**

> **Cooperate with independent researchers and civil society through funding and programs aimed at conducting evidence-based research on their services.**

> Share non-private metadata such as de-identified data or aggregate traffic statistics with independent researchers, in order to allow them to study these services and the movement of information without any identifying information.

318 See for instance
Garimella, K. & G. Tyson. (2018). WhatsApp, Doc? A First Look at WhatsApp Public Group Data. Retrieved from https://arxiv.org/abs/1804.01473 (Accessed on 20 August 2020).
Moreno, A.., P. Garrison, & K. Bhat. (2017). Whatsapp for Monitoring and Response During Critical Events: Aggie in the Ghana 2016 Election. Retrieved from https://collections.unu.edu/eserv/UNU:6189/iscram_unu_cs.pdf (Accessed on 20 August 2020).

319 See for instance
Pasquetto, I. V. et al. (2020). Understanding Misinformation on Mobile Instant Messengers (MIMs) in Developing Countries. Shorenstein Center. Retrieved from https://shorensteincenter.org/misinformation-on-mims/ (Accessed on 30 September 2020).

320 Global Partners Digital contribution to the working group on infodemics.

> Share non-private metadata through closed virtual environments[321] or research safe harbors,[322] in which researchers can access and analyze sensitive data (see part 2 of Chapter 1 on the audit of transparency requirements).

> Invest in more research into the implementation of differential privacy—a standard to anonymize individual user data while ensuring privacy—in order to share metadata with academic institutions in a private and secure way (see part 1.1 of Chapter 1).[323]

> Invest in more research in order to understand how their services are used in local contexts. This would allow service providers to be more proactive than reactive when faced with abuses of the use of their services.

RECOMMENDATIONS TO **ACADEMIC COMMUNITY**

> Develop ethical standards—or adapt existing ones—that enable research to be carried out on closed groups within messenger services.[324]

> Research the implication of mechanisms and protocols (such as the 'franking mechanism'—see part 5.2 of this chapter) implemented by companies on the security of end-to-end encrypted messaging systems.

321 AlgorithmWatch contribution to the working group on infodemics, laying out a comprehensive data-access framework for platforms.
322 Aral, S. (2020). *The Hype Machine*. London: Random House. p.312.
323 Aral, S. (2020). *op. cit.*, p.216.
It defines differential privacy as 'a standard for making individuals' data anonymous, so that it can be examined to understand patterns of election manipulation and crime, while guaranteeing individual consumers' anonymity'.
324 Riedel, A. C. (2020). Behind Closed Curtains, Disinformation on Messaging Services.
Retrieved from https://shop.freiheit.org/#!/Publikation/918 (Accessed on 5 October 2020). p. 22.

NEXT STEPS

A new global governance structure for digital technology is needed to ensure the effective and coordinated democratic oversight of platforms. The recommendations of this working group could offer inspiration for its initial framework.

'Disinformation'[325] and 'misinformation'[326] are endemic to the online service providers that structure the global information and communication space[327], as their business model relies on data collection and emotional engagement to capture users' attention. The online service providers monetize advertisements and user-targeting. This business model has been called by some 'a disinformation for profit business model'.[328]

Disinformation on a large scale could be considered as a flipside to the coin of the marvels produced by the digital platforms. The MIT Professor Sinan Aral[329] found that false news stories on Twitter spread six times faster than true ones,[330] and reached 100,000 people on average, compared to 1,000. 'False stories diffused further, faster, deeper, and more broadly than the truth, in every category of information that we studied,' says Aral.[331] 'Sometimes by an order of magnitude.'

It is time to supplement the current self-regulation framework that the private companies benefit from. It is time to proportionally and intelligently regulate and co-regulate,[332] and even to consider outlawing some current practices in order to protect the users and our democracies. It is time to consider users' safety and our democracies' safety as the priority.

The quasi-daily announcements of new content moderation policies by Facebook, Twitter and other online service providers are simply firefighting measures, reinforcing the urgency to create a new regulation model. The fire is not contained.

Companies should apply precautionary principles. Some even consider that platforms should have a fiduciary duty towards users, and that some ought to be regulated as natural monopolies, public utilities or essential service facilities carriers.

This working group looked at four structural challenges, however there is an urgent need to address interconnected issues such as providing more funding to independent journalism[333] and independent academic research on the impact platforms make, expanding media and information literacy,[334] and fact-checking.

---

325 Disinformation: Information that is false and deliberately created to harm a person, social group, organization or country, as defined by UNESCO:
UNESCO. (2020). *Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training.*
Retrieved from https://en.unesco.org/fightfakenews (Accessed on 21 October 2020).

326 Misinformation: Information that is false but not created with the intention of causing harm, as defined by UNESCO, *op. cit.*

327 As defined in the Partnership on Information and Democracy:
International Partnership on Information & Democracy (2019).
Retrieved from https://informationdemocracy.org/international-partnership-on-information-democracy/ (Accessed on 21 October 2020).

328 By some experts interviewed in the Netflix documentary, *The Social Dilemma.*

329 Member of the Steering Committee.

330 Blanding, M. (2020). How Hype Proliferates. *Spectrum.*
Retrieved from https://spectrum.mit.edu/spring-2020/how-hype-proliferates/ (Accessed on 5 October 2020).

331 Aral's most recent book is *The Hype Machine*, published in September 2020 by Random House.

332 State regulation, enforced by governments; self-regulation, exercised by platforms; and co–regulation, undertaken by governments and platforms together through mandatory or voluntary agreements, as summarized in
Pírková, E. & J. Pallero. (2020). *26 Recommendations on Content Governance: a Guide for Lawmakers, Regulators, and Company Policy Makers.* Access Now. Retrieved from https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (Accessed on 21 October 2020).

333 EURACTIV (2020). Media Recovery Beyond Regulation: a NEWS Bundle Across Creative Europe and Horizon, like MEDIA.
Retrieved from https://www.euractiv.com/section/digital/opinion/media-recovery-beyond-regulation-a-news-bundle-across-creative-europe-and-horizon-like-media/ (Accessed on 21 October 2020).

334 Posetti, J. & K. Bontcheva. (2020). *Freedom of Expression and Addressing Disinformation on the Internet.*
Chapter 8 : Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Broadband Commission.
Retrieved from https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf (Accessed on 19 October 2020).

Even more importantly, there is an urgency to address competition issues, the question of a platform's liability,[335] and real data protection.[336] Data is the currency of the digital world.[337] As the former chairman of the Federal Communications Commission (FCC), Tom Wheeler put it: "It's time to focus on the data hoarding by the platforms and not its effects."[338] It would probably make sense in the short term that the Forum on Information and Democracy creates a specific working group focused on data protection in the global context.[339]

New approaches to safety and quality standards in design should be explored further. Closed messaging services create complex challenges in the fight against disinformation. End-to-end encryption, secrecy of correspondence, and people's right to privacy need to be protected in these new mixed private-public spaces. Many answers still need to be found.

For all these interconnected issues, regulation and co-regulation should continue to be explored carefully, and should refrain from exclusively following a content-takedown approach, which could lead to creating a censorship framework. Innovation should not be stifled by regulation. Simultaneously, preventing a shift of states' obligations onto platforms, a privatization of law enforcement, or speech regulation is crucial. Current regulation attempts by authoritarian regimes and consolidated democracies in this respect are concerning. 'A democratic and balanced regulatory system should also protect platforms from the illegitimate pressures of governments and other stakeholders,' Observacom rightly points out in their contribution to this working group.[340]

This Forum on Information and Democracy can play a leading role in inventing new models of public regulation and co-regulation[341] for the much-needed global governance framework.[342]

This has been done for previous technological revolutions, such as with the radio, airplanes, pharmaceuticals or nuclear energy. One of the new challenges here is to adapt to the ever-changing nature of artificial intelligence (AI) and machine-learning algorithms. This is possible.

The steps towards concrete change will be led by a group of like-minded governments working closely with the civil society, such as the states who endorsed the international Partnership on Information and Democracy.

These recommendations should guide them in creating the new global governance structure for digital technology.

---

335 There is currently no corporate liability under international law for platforms whose products, services, or operations cause, contribute to, or are directly linked to the commission of human rights violations. Meanwhile, various states have initiated or are presently exploring the imposition of liability on platforms for content moderation.

336 Interesting latest developments with the California Consumer Privacy Act, the Global Privacy Control, etc.

337 'The currency extracted from individuals in the consumer internet context is typically not money, but rather a novel, complex combination of individuals' personal data and attention.' See Ghosh, D. (2020). *Terms of Disservice.* Washington, DC: Brookings. p. 216.

338 During panel discussion Breaking Up Big Tech, organized by the Brookings Institution on 4 August 2020. See, https://www.brookings.edu/events/breaking-up-big-tech/.

339 Exploring limitation of data collection, harmful microtargeting, regulation of data brokerage, digital influence industry and enforcement of data portability.

340 Observacom et. al., (July 2020). Standards for the Democratic Regulation of Large Content Platforms to Ensure Freedom of Expression Online and an Open and Free Internet. Contribution to this working group.

341 Such as a Digital Stability Board (See: https://www.cigionline.org/articles/digital-platforms-require-global-governance-framework), Social Media Council(s) (See: https://www.cigionline.org/articles/social-media-council-bringing-human-rights-standards-content-moderation-social-media) and Internet Ombudsman (See: https://pace.coe.int/en/files/28728/html.)

342 Co-regulation refers to a system in which the general guidelines and expected results of platform policies are defined in a legal instrument, with input from multiple sectors, which must be applied directly by platforms taking into consideration local and regional context and in line with human rights principles. An appropriate body, with guarantees of independence and autonomy, should oversee the companies' application of these standards. Observacom et. al., (July 2020), *op. cit.*

SELECTED **BIBLIOGRAPHY**

CHAPTER 1

Annenberg Public Policy Center of the University of Pennsylvania (2020). *Freedom and Accountability A Transatlantic Framework for Moderating Speech Online.* Retrieved from: https://www.annenbergpublicpolicycenter.org/feature/transatlantic-working-group-freedom-and-accountability/ (Accessed on 3 August 2020).

Bradford, B., F. Grisel, T. Meare, E. Owens, B. Pineda, J. Shapiro, T. Tyler, & D. Peterman. (2019). *Report Of The Facebook Data Transparency Advisory Group*. The Justice Collaboratory. Retrieved from https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf (Accessed on 5 September 2020).

Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision. (2019) . Retrieved from: https://minefi.hosting.augure.com/Augure_Minefi/r/ContenuEnLigne/Download?id=AE5B7ED5-2385-4749-9CE8-E4E1B36873E4&filename=Mission%20Re%CC%81gulation%20des%20re%CC%81seaux%20sociaux%20-ENG.pdf (Accessed on 20 September 2020).

European Partnership for Democracy. (March 2020). Virtual Insanity? The Need to Guarantee Transparency in Digital Political Advertising. Retrieved from https://epd.eu/wp-content/uploads/2020/04/Virtual-Insanity-synthesis-of-findings-on-digital-political-advertising-EPD-03-2020.pdf (Accessed on 6 September 2020).

Kaye, D. (2018). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Retrieved from https://www.undocs.org/A/HRC/38/35 (Accessed on 10 August 2020).

Pírková, E. & J. Pallero. (2020). *26 Recommendations on Content Governance: a Guide for Lawmakers, Regulators, and Company Policy Makers.* Access Now. Retrieved from https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (Accessed on 21 October 2020).

MacCarthy, M. (2020). *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*. Annenberg Public Policy Center. Retrieved from https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Transparency_TWG_MacCarthy_Feb_2020.pdf (Accessed on 18 August 2020).

MacCarthy, M. (2020). A Dispute Resolution Program for Social Media Companies. Brookings. Retrieved from https://www.brookings.edu/research/a-dispute-resolution-program-for-social-media-companies/ (Accessed on 20 October 2020).

Maréchal, N. & E. Roberts Biddle. (2020). It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge. Key Transparency Recommendations for Content Shaping and Moderation. *Ranking Digital Rights*. Retrieved from https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/key-transparency-recommendations-for-content-shaping-and-moderation/ (Accessed on 2 September 2020).

Maréchal, N., R. MacKinnon, & J. Dheere. (2020). Getting to the Source of Infodemics: It's the Business Model. Key Recommendations for Policymakers. *New America*. Retrieved from https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model/key-recommendations-for-policymakers (Accessed on 1 September 2020).

Observacom. (July 2020). A Latin American Perspective for Content Moderation Processes that are Compatible with International Human Rights Standards—Contribution to this working group.

Posetti, J. & K. Bontcheva. (2020). *Freedom of Expression and Addressing Disinformation on the Internet*. Chapter 8 : Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Broadband Commission. Retrieved from https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf (Accessed on 19 October 2020).

Wheeler, T., P. Verveer, & G. Kimmelman. (2020). New Digital Realities; New Oversight Solutions. Retrieved from https://shorensteincenter.org/new-digital-realities-tom-wheeler-phil-verveer-gene-kimmelman/ (Accessed on September 1, 2020).

CHAPTER 2

Barrett, Paul M. (June 2020). *Who Moderates the Social Media Giants? A Call to End Outsourcing*. NYU Stern Center for Business and Human Rights. Retrieved from https://bhr.stern.nyu.edu/tech-content-moderation-june-2020. (Accessed on October 25, 2020).

Council of Europe Parliamentary Assembly. (2020). Towards an Internet Ombudsman Institution. Retrieved from https://pace.coe.int/en/files/28728/html (Accessed on 12 October 2020).

Domino, J. (2 January 2020). How Myanmar's Incitement Landscape can Inform Platform Regulation in Situations of Mass Atrocity. *Opinio Juris*. Retrieved from http://opiniojuris.org/2020/01/02/how-myanmars-incitement-landscape-can-inform-platform-regulation-in-situations-of-mass-atrocity/ (Accessed on October 12, 2020).

Douek, E. (2020). COVID-19 and Social Media Content Moderation. *Lawfare*. Retrieved from https://www.lawfareblog.com/covid-19-and-social-media-content-moderation (Accessed on 12 October 2020).

Fay, R. (2019). Digital Platforms Require a Global Governance Framework (on Creating a Digital Stability Board). Center for International Governance Innovation. Retrieved from https://www.cigionline.org/articles/digital-platforms-require-global-governance-framework (Accessed on 12 October 2020).

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Retrieved from: https://yalebooks.yale.edu/book/9780300173130/custodians-internet

Gonzales, G. (2020). Increasing Sharing Friction, Trust, and Safety Spending may be Key Facebook Fixes. Rappler. Retrieved from https://www.rappler.com/technology/features/tristan-harris-aza-raskin-maria-ressa-undivided-attention-podcast (Accessed on 1 November 2020).

Kaye, D., (2015). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Retrieved from https://freedex.org/wp-content/blogs.dir/2015/files/2015/10/Dkaye_encryption_annual_report.pdf (Accessed 17 August 2020).

Kaye, D. (2019). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Retrieved from https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf (Accessed on 17 August 2020).

Observacom. (2020). A Latin American Perspective for Content Moderation Processes that are Compatible with International Human Rights Standards—Contribution to this working group.

Ong, J. C. & R.Tapsell. (May 2020). *Mitigating Disinformation in Southeast Asian Elections: Lessons from Indonesia, Philippines and Thailand*. NATO StratCom Centre of Excellence. Retrieved from www.stratcomcoe.org/mitigating-disinformation-southeast-asian-elections (Accessed on 12 October 2020).

Organization for Security and Cooperation in Europe (OSCE). (2020). *Joint Declaration on Freedom of Expression and Elections in the Digital Age*. Retrieved from: https://www.osce.org/representative-on-freedom-of-media/451150 (Accessed on 2 November 2020).

Marantz, A. (2020). Why Facebook Can't Fix Itself. *The New Yorker.* Retrieved from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself?mc_cid=a2705e31cc&mc_eid=dd9bd17d22 (Accessed on 20 October 2020).

Scheinin, M. (2010). *Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism. Ten Areas of Best Practices in Countering Terrorism*. A/HRC/16/51 Human Rights Council.
Retrieved from https://documents-dds-ny.un.org/doc/UNDOC/GEN/G10/178/98/PDF/G1017898.pdf?OpenElement

Schiffrin, A. (2020). *Beyond Transparency: Regulating Online Political Advertising*. Roosevelt Institute.
Retrieved from https://rooseveltinstitute.org/publications/beyond-transparency-regulating-online-political-advertising/ (Accessed on 27 October 2020).

Wingfield, R., I. Tuta, & T. Bansal, (2020). *The Tech Sector and National Action Plans on Business and Human Rights*. The Danish Institute of Human Rights. Retrieved from https://www.humanrights.dk/publications/tech-sector-national-action-plans-business-human-rights (Accessed on 12 October 2020).

United Nations. (2011). The UN Guiding Principles on Business and Human Rights: Implementing the 'Protect, Respect, Remedy' Framework was Unanimously Endorsed by the UN Human Rights Council in Resolution 17/4.
Retrieved from https://undocs.org/en/A/HRC/RES/17/4 (Accessed on 12 October 2020).

United Nations. (2019). *UN Strategy and Plan of Action on Hate Speech.* Retrieved from: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf (Accessed on 1 November 2020).

United Nations. (2020). *United Nations Guidance Note on Addressing and Countering COVID-19-Related Hate Speech.*
Retrieved from: https://www.un.org/en/genocideprevention/documents/Guidance%20on%20COVID-19%20related%20Hate%20Speech.pdf (Accessed on 1 November 2020).

United Nations Human Rights Office of the High Commissioner. (2020). *Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence.* Retrieved from: https://www.ohchr.org/EN/NewsEvents/Pages/TheRabatPlanofAction.aspx (Accessed on 1 November 2020).

CHAPTER 3

Goodman, E. (2020). Digital Information Fidelity and Friction. Knight First Amendment Institute at Columbia University. Retrieved from https://knightcolumbia.org/content/digital-fidelity-and-friction (Accessed on 8 October 2020).

Posetti, J. & K. Bontcheva. (2020). *Freedom of Expression and Addressing Disinformation on the Internet*.
Chapter 8 : Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Broadband Commission.
Retrieved from https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf (Accessed on 9 October 2020).

Radsch, Courtney C. (2020). Tech Platforms Struggle to Label State-controlled Media. Committee to Protect Journalists. Retrieved from https://cpj.org/2020/08/tech-platforms-struggle-to-label-state-controlled-media/ (Accessed on 2 October 2020).

Simpson, E., & A. Conner. (2020). Fighting Coronavirus Misinformation and Disinformation. Center For American Progress. Retrieved from https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation/ (Accessed on 8 October 2020).

Steenfadt, O. (2020). Sustaining Journalism During COVID-19, How the EU Can Turn Digital Platform Regulation Into a Tool for Democracy. *Friedrich-Ebert-Stiftung*. Retrieved from http://library.fes.de/pdf-files/bueros/budapest/16406.pdf (Accessed on 2 October 2020).

Steenfadt, O., E. Mazzoli, & S. Luca. (2020). Ensuring The Visibility and Sustainability of Reliable, Accurate Information in the Online Sphere Through Voluntary Standards – A Co-Regulatory Approach (Accessed on 2 October 2020).

Wylie, C. (2019). *Mindf*ck – Cambridge Analytica and the Plot to Break America.* New York: Random House.

CHAPTER 4

Barrett, P. M. (2019). *Disinformation and the 2020 Election: How the Social Media Industry Should Prepare*. NYU Stern Center. Retrieved from https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_election_2020_report?fr=sY2QzYzI0MjMwMA (Accessed on 29 September 2020).

Bowles J., H. Larreguy, & S. Liu. (2020). *Countering Misinformation via WhatsApp: Evidence From the COVID-19 Pandemic in Zimbabwe.* Center for International Development at Harvard University. Retrieved from https://www.hks.harvard.edu/sites/default/files/centers/cid/files/publications/faculty-working-papers/2020-05-CID-WP-380-Countering%20Misinformatoin%20Via%20WhatsApp-Evidence%20from%20the%20COVID-19%20Pandemic%20in%20Zimbabwe.pdf (Accessed on 3 September 2020).

Coca, N. (2020). Disinformation from China Floods Taiwan's Most Popular Messaging App. *Coda Story.* Retrieved from https://www.codastory.com/authoritarian-tech/taiwans-messaging-app/ (Accessed on 13 October 2020).

De Freitas Melo, P., C. Coimbra Vieira, K. Garimella, P. O. S.Vaz de Melo, & F. Benevenuto. (2019). Can WhatsApp Counter Misinformation by Limiting Message Forwarding. Retrieved from https://arxiv.org/pdf/1909.08740.pdf (Accessed on 23 September 2020).

Global Partners Digital. (2017). Encryption Policy for Human Rights Defenders. Retrieved from https://www.gp-digital.org/wp-content/uploads/2017/09/TRAVELGUIDETOENCRYPTIONPOLICY.pdf (Accessed on 21 August 2020).

Kaye, D., (2015). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Retrieved from https://freedex.org/wp-content/blogs.dir/2015/files/2015/10/Dkaye_encryption_annual_report.pdf (Accessed 17 August 2020).

Kaye, D. (2019). *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*. Retrieved from https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf (Accessed on 17 August 2020).

Knuutila, A.., A. Herasimenka, J. Bright, R. Nielsen, & P. N. Howard. (2020). Junk News Distribution on Telegram. Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/07/Junk-News-Distribution-on-Telegram.-Data-Memo.pdf (Accessed on 23 September 2020).

Pírková, E. & J. Pallero. (2020). *26 Recommendations on Content Governance: a Guide for Lawmakers, Regulators, and Company Policy Makers.* Access Now. Retrieved from https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf (Accessed on 21 October 2020).

Portnoy, E. (2019). Why Adding Client-Side Scanning Breaks End-To-End Encryption. Electronic Frontier Foundation. Retrieved from https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption (Accessed on 18 September 2020).

Reis, J. C. S., P. Melo, K. Garimella, F. Benevenuto. (2020). Can WhatsApp Benefit from Debunked Fact-Checked Stories to Reduce Misinformation?. *The Harvard Kennedy School (HKS) Misinformation Review*. Retrieved from https://doi.org/10.37016/mr-2020-035 (Accessed on 23 August 2020).

Riedel, A. C. (2020). Behind Closed Curtains, Disinformation on Messaging Services. Retrieved from https://shop.freiheit.org/#!/Publikation/918 (Accessed on 5 October 2020).

Rodriguez, K. & S. Schoen. (2020). FAQ: Why Brazil's Plan to Mandate Traceability in Private Messaging Apps Will Break User's Expectation of Privacy and Security. Retrieved from https://www.eff.org/fr/deeplinks/2020/08/faq-why-brazils-plan-mandate-traceability-private-messaging-apps-will-break-users (Accessed on 18 September 2020).

ACKNOWLEDGMENTS

The Forum would like to warmly thank the Steering Committee members of this working group, especially its co-chairs, and the more than 60 experts we interviewed and the 30 organizations and experts who prepared written contributions, for their time, insightful perspective and support.

AMONG THEM:

- Susan Juliet Agwang, *Africa Freedom of Information Centre*
- Kamel Ajji, *doctoral student, Paris 2 Panthéon-Assas University*
- Susie Alegre, *associate tenant, Doughty Street Chambers*
- Veridiana Alimonti, *Latin American senior policy analyst, Electronic Frontier Foundation*
- Ian Barber, *legal officer, Global Partners Digital*
- Paul Barrett, *deputy director, NYU Stern Center for Business and Human Rights*
- Chris Bealls, *policy lead, CIGI*
- Susan Benesch, *executive director, Dangerous Speech Project*
- Priyanjana Bengani, *senior research fellow, Columbia Journalism School's Tow Center for Digital Journalism*
- Max Beverton-Palmer, *head of internet policy, Tony Blair Institute*
- Constance Bommelaer, *Area Vice President – Institutional Relations, The Internet Society*
- Charles Bradley, *executive director, Global Partners Digital*
- Emma Briant, *researcher, Bard College*
- Nadia Cabral, *advocacy and media consultant, Avaaz*
- Scott Campbell, *professor of communication and media, University of Michigan*
- Lucien Castex, *co-chair, French Internet Governance Forum*
- Julie E. Cohen, *Georgetown University Law Center*
- Paul Coppin, *head of legal desk, Reporters Without Borders (RSF)*
- Noemi Dado, *editorial board member, BlogWatch*
- Emma Daly, *communications director, Human Rights Watch*
- Camille Darche, *doctoral student, University Paris-Nanterre*
- Giovanni De Gregorio, *PhD candidate in public law, University of Milano-Bicocca*
- Agustina Del Campo, *director, Center for Studies on Freedom of Expression and Access to Information (CELE) – Universidad de Palermo*
- Emmanuel Didier, *research director, CNRS*
- Tidiane Dioh, *responsible for the support program for the media and journalists associations, Organisation internationale de la Francophonie*
- Pierre François Docquir, *ARTICLE 19*
- Cécile Dolbeau, *researcher, Caen University*
- Evelyn Douek, *Harvard Law School*
- Robert Fay, *managing director of digital economy, CIGI*
- Frederic Filloux, *entrepreneur, editor of mondaynote.com*
- Kiran Garimella, *postdoctoral researcher, MIT Institute for Data, Systems, and Society*
- Akriti Gaur, *independent consultant, India*
- Dipayan Ghosh, *co-director of the Platform Accountability Project, Shorenstein Center on Media, Politics and Public Policy – Harvard Kennedy School*
- Gustavo Gomez, *director, Observacom*
- David Greene, *Civil Liberties director, EFF*
- Joao Brant, *senior researcher, Observacom*
- Tonei Glavinic, *director of operations, Dangerous Speech Project*
- Tristan Harris, *president, Center for Humane Technology*
- Jonathan Hew, *lawyer*
- Dunstan Allison-Hope, *BSR*
- Jamie Joyce, *executive director, the Society Library*
- Argyro Karanasiou, *director, LETS Lab*

- Roukaya Kasenally, *CEO, African Media Initiative*
- Usama Khilji, *director, Bolo Bhi*
- Claude Kirchner, *director, National Pilot Committee for Digital Ethics*
- Anunay Kulshrestha, *doctoral student, Princeton University*
- Tawana Kupe, *Vice-Chancellor, University of Pretoria*
- Ozan Kuru, *assistant professor, National University of Singapore*
- Horacio Larreguy, *associate professor of government, Harvard University*
- Théophile Lenoir, *head of digital program, Institut Montaigne*
- Gabrielle Lim, *Harvard Kennedy School's Shorenstein Center*
- Benoît Loutrel, *inspector general, INSEE and former director, ARCEP*
- Mark MacCarthy, *senior fellow, Institute for Technology Law and Policy – Georgetown Law*
- Megan MacDuffee Metzger, *associate director for research at the Global Policy Incubator, Stanford*
- Dani Madrid-Morales, *assistant professor of journalism, Jack J. Valenti School of Communication – University of Houston*
- Pere Masip, *faculty member, Blanquerna Faculty of Communication – Ramon Llull University*
- Anna Mazgal, *EU policy advisor, Wikimedia*
- Nathan Miller, *campaign director, Avaaz*
- Blaise Ndola, *program manager, ICT programs at Rudi International*
- Rasmus Nielsen, *director, Reuters Institute for the Study of Journalism*
- Onora O'Neill, *professor of philosophy, University of Cambridge*
- Aviv Ovadya, *founder, the Thoughtful Technology Project*
- Javier Pallero, p*olicy director, Access Now*
- Smriti Parsheera, *policy researcher, National Institute of Public Finance and Policy*
- Torry Pedersen, *head of editorial, Schibsted*
- Jan Penfrat, *senior policy advisor, European Digital Rights*
- Eliska Pirkova, *Europe policy analyst, Access Now*
- Andres Plazza, *Observacom*
- Jack Poulson, *former research scientist, Google*
- Ruth Price, *project lead, Digital Action*
- Courtney Radsch, *advocacy director, Committee to Protect Journalists*
- Friederike Reinhold, *Senior Policy Advisor, Algorithm Watch*
- Katitza Rodriguez, *international rights director, Electronic Frontier Foundation*
- Gilbert Sendugwa, *Africa Freedom of Information Centre*
- Sonja Solomun, *Research Director, Centre for Media, Technology and Democracy, McGill*
- Antonia Staats, *senior campaigner, Avaaz*
- Alice Stollmeyer, *executive director, Defend Democracy*
- Ross Tapsell, *Australian National University*
- Heidi Tworek, *assistant professor of history, University of British Columbia*
- José Van Dijck, *professor of media and digital society, Utrecht University*
- Serena Villata, member, *National Pilot Committee for Digital Ethics*
- Ben Wagner, *director, Privacy & Sustainable Computing Lab – Vienna University of Economics and Business*
- Maeve Walsh, *Carnegie UK Trust*
- Jeremy West, *senior policy analyst, Directorate for Science, Technology and Innovation – OECD*
- Richard Wingfield, *head of legal, Global Partners Digital*
- Lorna Woods, *professor of internet law, University of Essex*
- Narasimha Sai Yamanoor, *senior IoT applications engineer, Linde*
- Srihari Yamanoor, *senior manufacturing engineer, Think Surgical*
- Jillian York, *director for international freedom of expression, Electronic Frontier Foundation*
- Nicolo Zingales, *professor of information law and regulation, Fundação Getulio Vargas*
- Célia Zolynski, *member, National Pilot Committee for Digital Ethics*

INTERNATIONAL PARTNERSHIP ON INFORMATION AND DEMOCRACY

1. We, the States taking part in the International Partnership for Information and Democracy;

2. Recalling the right to freedom of opinion and expression that includes freedom to hold opinions without interference and the right to seek, receive and impart information and ideas through any media and regardless of frontiers;

3. Recognizing the rapid evolution of the global information and communication space, including the development of internet;

4. Acknowledging that the global information and communication space is a shared public good of significant democratic values that requires special protection to ensure it remain global, open and accessible to all, and that the actions of public authorities or private actors should not unduly restrict this space directly or indirectly;

5. Underlining that this new global information and communication space has enhanced the possibilities of exercising the right to freedom of opinion and expression and improved access to information in many ways;

6. Recognizing that, however, it is not immune from new ways to censor, manipulate and control information;

7. Reiterating our commitment to protect all human rights, including the right to freedom of opinion and expression, guaranteed by Article 19 of both the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights;

8. Welcoming multi-stakeholder efforts to establish, at international level, "a people-centred, inclusive and development-oriented information society, where everyone can create, access, utilize and share information and knowledge", as pledged at the World Summit on the Information Society adopted on 12 December 2003;

9. Taking note of all relevant UN resolutions and conventions related to safety of journalists and the promotion, protection and enjoyment of human rights on the Internet and recalling in particular the 2005 UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions' guiding principle that cultural diversity can be protected and promoted only if human rights and fundamental freedoms, such as the right to freedom of opinion and expression are guaranteed;

10. Welcoming the adoption of the 2030 Agenda for Sustainable Development and the commitments therein to, inter alia, promote peaceful and inclusive societies for sustainable development, including by ensuring public access to information and protecting fundamental freedoms, in accordance with national legislation and international agreements, and therefore recognizing the important contribution of the promotion and protection of the safety of journalists in this regard;

11. Recalling the International Declaration on Information and Democracy adopted on 5 November 2018 by the independent international Information and Democracy Commission, initiated by the non-governmental organization Reporters Without Borders (RSF);

12. Considering the Joint Statement made by 12 Heads of State and Government on 11 November 2018 at the Paris Peace Forum announcing their desire to launch an initiative inspired by the work of this Commission;

13. Underlining that information can be regarded as reliable insofar as its collection, processing and dissemination are free, independent, diverse and based on cross-checking of various sources, in a pluralistic media landscape where the facts can give rise to a diversity of interpretation and viewpoints;

14. Considering that the right to freedom of opinion and expression is essential for the enjoyment of other human rights and fundamental freedoms and that access to reliable information is crucial to the exercise of freedom of opinion;

15. Commending as such the crucial role of journalism for freedom of opinion and expression and to contribute to and foster debate over issues of public interest both online and offline;

16. Reaffirming that independent media are essential to a free and open society and accountable systems of government and are of particular importance in safeguarding human rights and fundamental freedoms;

17. Expressing concern at the damage which can be done by the spreading of false or manipulated information intended to deliberately deceive, and recognising the role that the collection, processing and dissemination of information, when free, independent, diverse and based on cross-checking of variable sources, can play in mitigating that damage;

18. Highlighting the importance of transparency surrounding media ownership, financing and editorial independence;

19. Observing that access to the global information and communication space is of crucial importance for full democratic participation;

20. Underlining in this regard the importance of education to media and information for people to be free, critical, independent, and capable of defending oneself against misinformation, disinformation and manipulation of public opinion;

21. Recognizing also the importance of public trust in and the credibility of journalism, and the challenges it faces in maintaining journalistic professionalism in an environment where targeted disinformation and smear campaigns to discredit the work of journalists are increasing;

22. Noting with concern that all forms of human rights violations and abuses committed against journalists which affect directly their safety and prevent them from providing information to the public negatively affect the exercise of the right to freedom of expression;

23. Welcoming a multi-stakeholder approach to internet governance;

24. Confirming that all stakeholders, especially the Online Service Providers, as they help structure the information and communication space by creating technical means, architectures and standards for information and communication, have responsibilities relating to their role;

25. Affirm the following principles:

a. The global information and communication space, which is a shared public good of significant democratic value, must support the exercise of human rights, most notably the right to freedom of opinion and expression, including the freedom to seek, receive and impart information and ideas of all kinds, through any media of one's choice regardless of frontiers, in accordance with the International Covenant on Civil and Political Rights (Article 19);

b. Access to reliable information must be protected and promoted to enable democratic participation and the exercise of freedom of opinion and expression;

c. Information can be regarded as reliable insofar as its collection, processing and dissemination are free and independent, based on cross-checking of various sources, in a pluralistic media landscape where the facts can give rise to a diversity of interpretation and viewpoints;

d. In accordance with the international law and standards on the right to freedom of opinion and expression, journalists and media workers, in the course of their function, must be protected against all forms of violence, threats, and discrimination; against all forms of arbitrary detention, abusive legal proceedings; against any unduly restrictive efforts to prevent them from carrying out their works and have access to appropriate legal remedies, including as relevant with respect to the confidentiality of their sources;

e. Sustainable business models must be developed to serve high-quality independent journalism;

26. Call on the online service providers that structure global information and communication space to:

a. Comply with the principles of transparency, accountability, and political, ideological and religious neutrality, including with regard to their own services, while bearing in mind their responsibilities in this matter, and implement mechanisms to foster access to reliable information and to counter the dissemination of false or manipulative information intended to deceive audiences;
b. Uphold the responsibilities incumbent on them according among others to the UN principles on business and human rights ahead of the design of new programmes, software and connected devices;
c. Demonstrate transparency and accountability in algorithmic curation, including moderating human and technical decision-making processes, financial promotion of online content, collection of personal data and relevant agreements concluded with any government or private entity which have an impact on compliance with the above principles;
d. Ensure the consistency of their policies, procedures, algorithmic design, and tools for the moderation and curation of content with human rights and in particular international standards on the right to freedom of opinion and expression;
e. Enable access to a variety of media outlets, information and ideas through diverse indexing solutions that limit the risk of echo-chambers and filter bubbles that are algorithmically populated;
f. Promote tools to foster visibility and dissemination of reliable information;

27. We will strive to:
a. Implement international obligations related to the right to freedom of opinion and expression as well as media freedom, including by respecting, promoting and protecting the freedom to seek, receive and impart information regardless of frontiers;
b. Ensure that our legislations, policies and procedures promote a global space that fosters access to reliable information that complies with the above principles;
c. Promote national and international legal frameworks that are compliant and conducive to the aforementioned right to freedom of opinion and expression, and that are able to allocate clear obligations and responsibilities;
d. Encourage public awareness and exercise of this right;
e. Work to prevent acts of violence, threats and attacks aimed at journalists and media workers, and fight against impunity for crimes committed against journalists through the conduct of impartial, prompt, thorough, independent and effective investigations;
f. Condemn unequivocally and address the specific attacks on women journalists and media workers in the exercise of their work, including sexual and gender-based discrimination and violence, intimidation and harassment, online and offline;
g. Establish and preserve a safe environment enabling journalists and media workers to work freely and independently and without undue external interference and intimidation and without any form of discrimination;
h. Combat any unduly restrictive measure against the right to freedom of opinion and expression and take measures to prevent manipulation of information by state or non-state actors, and condemn, prevent and fight such actions;
i. Sustain and support conditions to ensure the financial viability of journalism, while ensuring that this support empowers and does not undermine editorial independence and journalistic freedom;
j. Promote this Partnership as a means to reinforce existing international and regional mechanisms contributing to the implementation of established human rights instruments and Sustainable Development Goal 16.10 ; to facilitate multi-stakeholder discussions on the means, standards and exchange experience ; and to encourage development of self-regulation practices among information and communication space actors;
k. Welcome the work carried out by Reporters Without Borders (RSF) to foster the implementation of this Partnership through the creation of a Forum in cooperation with other independent organisations, particularly to provide non-binding recommendations for States and Online service providers;
l. Convene in consultation with states signatories an annual meeting in an agreed upon format;
m. Promote the International Partnership for Information and Democracy among all States with a view to encouraging them to join it;

28. Membership of this Partnership is open to all States, after approval by the signatory States.

INTERNATIONAL DECLARATION
ON INFORMATION AND DEMOCRACY

PREAMBLE

The global communication and information space is a common good of humankind and should be protected as such. Its management is the responsibility of humankind in its entirety, through democratic institutions, with the aim of facilitating real communication between individuals, culture, peoples and nations, in the service of human rights, civil concord, peace, life and the environment.

The global communication and information space should serve the exercise of freedom of expression and opinion and shall respect the principles of pluralism, freedom, dignity, tolerance and the ideal of reason and understanding. Knowledge is necessary for human beings to develop their biological, psychological, social, political and economic capacities. Access to knowledge, particularly knowledge of reality, is a fundamental right.

Political control of the media, subjugation of news and information to private interests, the growing influence of corporate actors who escape democratic control, online mass disinformation, violence against reporters and editors, and the undermining of quality journalism, threaten the exercise of the right to knowledge. Any attempt to abusively limit it, whether by force, technology or legal means, is a violation of the right to freedom of opinion.

The communication and information space must be organized in such a way as to allow rights and democracy to be exercised. It should preserve and strengthen our ability to address challenges of the present time, to anticipate our common destiny and to help us shape global sustainable development which takes into account the rights and interests of future generations.

The communication and information space should guarantee the freedom, independence, and pluralism of news and information. As a common good, this space has social, cultural and democratic value and should not be reduced to its commercial dimension alone. Dominant positions in the production, distribution or curation of information, must be prevented where possible and controlled when unavoidable, in order to preserve the variety of facts and viewpoints.

PRINCIPLES

RIGHT TO INFORMATION

Freedom of opinion is guaranteed by the free exchange of ideas and information based on factual truths. The truth, which may take many forms, is grounded on the correspondence between reality and perceptions or on the best available evidence from established methods of scientific, academic, journalistic or other professional practices designed to produce trustworthy information and knowledge.

Reliable information underpins the exercise of freedom of opinion, respect for other human rights and all democratic practices, including deliberation, election, decision-making and accountability. The integrity of the democratic process is violated when information that could influence this process is manipulated.

The right to information consists of the freedom to seek, receive and access reliable information. Information can only be regarded as reliable when freely gathered, processed and disseminated, according to the principles of commitment to truth, plurality of viewpoints and rational methods of establishment and verification of facts.

The commitment to free pursuit of truth, factual accuracy and “do no-harm” principles is necessary for the integrity of news and information. Disseminating information that is misleading or incorrect or

withholding information that should be known can undermine the individuals ability to understand their environment and to develop their capacities.
Undisclosed conflicts of interest in the field of information pose a threat to freedom of opinion. Content that is designed to advertise or promote must be clearly identified as such.

FREEDOM OF EXPRESSION

Freedom of expression is a fundamental right of individuals to express themselves. In accordance with international standards on free speech and with due regard to the rights and reputation of others, it includes the right to criticize any system of thoughts and cannot be constrained or limited by the beliefs or sensitivities of others.

Intellectual property, which is only applicable to creations and inventions, should not create closed systems in the information and communication space and should not be used to restrict public deliberation. The product resulting from the creative work of gathering, processing and disseminating information confers the right to fair remuneration.

PRIVACY

Participants in the public debate must be able to protect the confidentiality of their private information or communications. The right to privacy may only be restricted, and only in a proportionate manner, where it is necessary in a democratic society for public order, the safety of persons, the prevention of crimes, the protection of health or the protection of the rights and freedoms of others.

RESPONSIBILITY

Responsibility of all participants in the public debate is a key principle, which implies transparency over their identity. Exceptions to the principle of transparency are legitimate if they facilitate the quest for truth or contribute to their own security.

All participants in the public debate are liable for their expression, including content they disseminate or help to disseminate. Liability may be established only on the basis of the restrictions on freedom of expression regarded as admissible under international standards.

TRANSPARENCY OF POWERS

Every public or private sector entity imbued with a form of power or influence has – within the limits of the public interest – transparency obligations in proportion to the power or influence it is able to exercise over people or ideas.

This transparency must be assured in a swift, sincere and systematic manner.

ENTITIES THAT CREATE MEANS, ARCHITECTURES OF CHOICE AND NORMS FOR INFORMATION AND COMMUNICATION

ACCOUNTABILITY

When creating technical means, architectures that shape choices and norms for communication, entities that contribute to the structure of the information and communication space shall respect the principles and guarantees that nourish and underpin the democratic nature of this space. They have to be held accountable in accordance with and in proportion to the impact of their contribution or participation.

POLITICAL, IDEOLOGICAL AND RELIGIOUS NEUTRALITY

These entities, such as platforms, shall comply fully with standards of freedom of expression and opinion and, to this end, shall respect political, ideological and religious neutrality when structuring the

information and communication space. Systems distributing or curating information and ideas must be neutral as regards the interests of those who control them, with the exception of advertising, which must be explicitly identified.

PLURALISM

Such entities, including platforms, shall promote diversity of ideas and information, media pluralism and favor serendipity. Tools used for curating and indexing information - meaning aggregating, sorting and prioritizing information - must provide alternative solutions, allowing for a pluralism of indexation, and allowing for freedom of choice for users.

RELIABLE INFORMATION

Such entities shall implement mechanisms that favor visibility of reliable information. Such mechanisms shall be based on criteria of transparency, editorial independence, use of verification methods and compliance with journalism ethics. The integrity, authenticity, traceability of ideas and information shall be promoted, so that their origin and mode of production and dissemination are known. It shall not be a violation of political, ideological and religious neutrality to favor reliable information.

TRANSPARENCY TO INSPECTION

Such entities must be predictable for those over whom they have influence, resistant to any manipulation and open to inspection. Platforms shall be transparent over curation algorithms, moderation (whether human or algorithmic), content sponsoring, collection of personal data, and agreements they may have entered into with governments.

INTEGRATION BY DESIGN

Compliance with the obligations of these entities, such as platforms, shall, as far as possible, be integrated from the outset into software, algorithms and connected objects. These corporate entities and services are required to observe due diligence.

MEDIA AND JOURNALISM

SOCIAL FUNCTION OF JOURNALISM

Journalism's social function is that of a "trusted third party" for societies and individuals. It allows for the establishment of checks and balances and empowers people to fully participate in society. It aims at giving account of reality, of revealing it in the broadest, deepest and most relevant manner possible, allowing for the exercise of the right to freedom of opinion.

Journalism's task is not just to portray events but also to explain complex situations and changes, being comprehensive and inclusive, allowing the public to distinguish the important from the trivial. It should reflect both positive and negative aspects of human activities and expose potential constructive solutions to important challenges.

JOURNALISM'S DEONTOLOGY

Journalists fulfil their social function when their rights are protected, when they can work freely and when they respect their professional obligations, as defined in the established ethical documents of the profession. Journalism can be practiced by a plurality of actors, without regard to their status, being professional or not.

Journalists must be committed to handling information in such a way as to serve the public interest and the public's fundamental rights. They should not treat information as a commodity. Motivated by the demands of truth, they must present the facts fairly, disregarding as much as possible their own interests and prejudices and rejecting all forms of connivance and conflicts of interest.

FREEDOM AND SAFETY OF JOURNALISTS

Journalism can only fulfill its social function if journalists' freedom and safety are guaranteed, online and offline. They must be protected against all forms of violence, pressure and discrimination, against all forms of abusive legal proceedings, and against any efforts to erode their ability to fulfill their social function.

They have the right to the protection of the confidentiality of their sources. An effective protection of whistleblowers is necessary in order to guarantee the transparency of powers.

EDITORIAL INDEPENDENCE

Journalists act in complete independence from all forms of power and undue influence, whether political, economic, religious or other. Any violation of the principles of independence, pluralism and honesty of information by public officials, owners, shareholders, advertisers or the media's commercial partners violates the freedom of information.

State or private sector funding for journalism should not be accompanied by conditions that would dictate the substance of content or seek to interfere with a journalist's professional judgement.

JOURNALISM SUSTAINABILITY

The social function of journalism justifies an effort by societies to ensure journalism financial sustainability.

TOWARD AN INTERNATIONAL FRAMEWORK FOR INFORMATION AND DEMOCRACY

Accountability for practices that cross diverse national boundaries raises complex challenges, particularly in a fast-changing field. Democratic accountability will require continuous expert participation that adequately balances global representation with rigorous evidence-based assessment of practices and conditions of knowledge production in the global communications and information space. To this end, an international group of experts should be created. Its funding and mandate shall provide sufficient independence from both companies and governments and it shall have the power to investigate practices and outcomes of the primary means, architectures, and norms of communications, on an ongoing basis, and issue periodic reports and recommendations on best practices.

INTERNATIONAL COMMISSION
ON INFORMATION AND DEMOCRACY

Initiated by Reporters Without Borders (RSF) in September 2018, the Commission published the International Declaration on Information & Democracy on 5 November 2018. It is composed of 25 people of 18 nationalities, including Nobel Laureates.

- Christophe Deloire, co-chair*, Secretary-General of Reporters Without Borders (RSF).*
- Shirin Ebadi, *co-chair, Founder of Defenders of Human Rights Centre and 2003 Nobel Peace Prize laureate.*
- Emily Bell, *Professor of Professional Practice at the Columbia University School of Journalism and director of the Tow Centre for Digital Journalism.*
- Yochai Benkler, *Faculty co-director of the Berkman-Klein Center for Internet & Society, Harvard University.*
- Teng Biao, *Academic lawyer and human rights activist, visiting scholar at the US-Asia Law Institute*.
- Nighat Dad, *Lawyer, internet activist, founder and executive director of the Digital Rights Foundation.*
- Primavera De Filippi, *Faculty associate at the Berkman-Klein Center for Internet & Society at Harvard University.*
- Mireille Delmas-Marty, *Emeritus professor at Collège de France and member of the Institut de France.*
- Abdou Diouf, *Former president of the Republic of Senegal and former secretary-general of the Organisation Internationale de la Francophonie (OIF).*
- Can Dündar, *Former editor-in-chief of the centre-left independent newspaper Cumhuriyet.*
- Francis Fukuyama, *Political scientist and political economist, professor at Stanford University.*
- Ulrik Haagerup, *Journalist and founder and CEO of Constructive Institute.*
- Hauwa Ibrahim, *Human rights lawyer and 2005 laureate of the European Parliament's Sakharov Prize.*
- Ann Marie Lipinski, *Curator of the Nieman Foundation for Journalism at Harvard University.*
- Adam Michnik, *Historian, journalist and essayist, editor-in-chief of Gazeta Wyborcza.*
- Eli Pariser, *Executive director at Upworthy, co-founder of Avaaz and chairman of MoveOn.*
- Antoine Petit, *Head of the French National Centre for Scientific Research (CNRS).*
- Navi Pillay, *Former UN High Commissioner for Human Rights and former president of the International Criminal Tribunal for Rwanda.*
- Maria Ressa, *Journalist and CEO of the Rappler news website.*
- Amartya Sen, *Economist, philosopher and 1998 Nobel laureate in Economic Sciences.*
- Joseph E. Stiglitz, *Economist, 1998 Nobel laureate in Economic Sciences.*
- Mario Vargas Llosa, *Writer, politician, journalist, college professor and 2010 Nobel laureate in Literature.*
- Marina Walker, *Journalist, deputy director of the International Consortium of Investigative Journalists.*
- Aidan White, *President & founder of the Ethical Journalism Network, former general secretary of the International Federation of Journalists.*
- Mikhail Zygar, *Founding editor-in-chief of the independent news TV-channel Dozhd, 2014 laureate of the International Press Freedom Award.*

BOARD AND STAFF OF THE FORUM ON INFORMATION AND DEMOCRACY

BOARD OF DIRECTORS

- Christophe Deloire, *secretary-general of Reporters without Borders (RSF), chair*
- Nighat Dad, e*xecutive director of the Digital Rights Foundation, vice-chair*
- Leon Willems, *director of Free Press Unlimited, treasurer*
- Bruce Girard, *senior researcher at Observacom*
- Sasha Havlicek, *founding CEO of the Institute for Strategic Dialogue*
- Alexa Koenig, *executive director of the Human Rights Center at Berkeley School of Law*
- Joe Powell, *deputy CEO of the Open Government Partnership*
- Aaron Shull, *general manager of the Center for International Governance Innovation*
- Henrik Urdal, *executive director the Peace Research Institute Oslo*
- Anri van der Spuy, *senior associate of the Research ICT Africa*
- Susan Wilding, *head of Geneva Office at Civicus*

PERMANENT SECRETARIAT*

- Camille Grenier, *project manager*
- Charlotte Caillat, *junior project officer*

* In addition to the Forum's team, the management of the Permanent Secretariat is currently delegated to Reporters without Borders (RSF).

RAPPORTEURS TEAM

- Delphine Halgand-Mishra, *executive director at the Signals Network, lead rapporteur*
- Iris de Villars, *head of tech desk, Reporters Without Borders (RSF)*
- Jenny Domino, *associate legal adviser, International Commission of Jurists (for Chapter 2: Meta-Regulation of Content Moderation)*

This working group was financially supported by the Délégation Société Civile of the Foreign Ministry of France, and the Gesellschaft für Zusammenarbeit (GIZ), the development agency of the German Republic.

Contact: contact@informationdemocracy.org

Forum on
Information
& Democracy