FUTURE LAW
WORKING PAPERS 2022 • 1

universität innsbruck
Institut für Theorie und Zukunft des Rechts

MATTHIAS C. KETTEMANN • MALTE KRAMME
CLARA RAUCHEGGER • CAROLINE VOITHOFER
EDITORS

A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation

JULIA HAAS
VIENNA

MAY 2022 • ZUKUNFTSRECHT@UIBK.AC.AT

FUTURE LAW WORKING PAPERS 2022 • 1

MATTHIAS C. KETTEMANN • MALTE KRAMME
CLARA RAUCHEGGER • CAROLINE VOITHOFER
EDITORS

universität innsbruck
Institut für Theorie und Zukunft des Rechts

The **Future Law Working Papers** was established in 2022 to offer a forum for cutting-edge research on legal topics connected to the challenges of the future. As the German Constitutional Court recently ruled, we have to act today to save the freedoms of tomorrow. Similarly, the Future Law Working Papers series hosts research that tackles difficult questions and provides challenging, and at times uncomfortable, answers, to the question of how to design good normative frameworks to ensure that rights and obligations are spread fairly within societies and between societies, in this generation and the next. The series is open for interdisciplinary papers with a normative twist and the editors encourage creative thinking. If you are interest in contributing, please send an email to the editors at **zukunftsrecht@uibk.ac.at.** Submissions are welcome in English and German.

The series is edited by the senior members of the Department of Legal Theory and the Future of Law at the University of Innsbruck, **Matthias C. Kettemann**, **Malte Kramme**, **Clara Rauchegger** and **Caroline Voithofer**.

Founded in 2019 as the tenth department of the law faculty, the **Department of Legal Theory and Future of Law** at the University of Innsbruck (ITZR) investigates how law can make individuals as well as society, states as well as Europe "fit" for the future and if and how law has to change in order to meet future challenges. This includes the preservation of freedom spaces as well as natural resources in an intergenerational perspective, the safeguarding of societal cohesion in times of technologically fueled value change, the normative framing of sustainable digitization and digitized sustainability, and the breaking through of traditional legal structures of domination and thought with a view to rediscovering the emancipatory element of law against law.

Publisher: Institut für Theorie und Zukunft des Rechts, Universität Innsbruck
Innrain 15, 6020 Innsbruck
Univ.-Prof. Mag. Dr. Matthias C. Kettemann, LL.M. (Harvard)
Univ.-Prof. Dr. Malte Kramme
Ass. Prof.in MMag.a Dr. Clara Rauchegger, LL.M. (Cambridge)
Univ.-Ass.in MMag.a Dr.in Caroline Voithofer

Suggested citation: Haas, A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation, Future Law Working Paper 2022-1, DOI: 10.25651/2022.01, https://doi.org/10.25651/2022.01

A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation

Mag.ª Julia Haas, LLM

Author

Julia Haas

@TheJuliaHaas

Julia is an international law and human rights expert. In her work, she has focused on international relations, the intersection of technology and human rights, and the prevention of marginalization. At the Office of the OSCE Representative on Freedom of the Media, Julia particularly focuses on policy research and development in the field of internet governance and digital participation, with a focus on the impact of artificial intelligence on freedom of expression, digital and legal safety of journalists, gender and online pluralism. Previously, she worked as legal officer and human rights adviser at the Austrian Ministry for Foreign Affairs. Julia sits on the advisory board for the Vienna Forum for Democracy and Human Rights, holds a Master's degree in Law from the University of Vienna, an Information and Media Law LLM and is a PhD candidate on the impact of digital innovations on freedom of the media.

Julia ist Expertin für internationales Recht und Menschenrechte. Ihr Fokus liegt auf internationalen Beziehungen, der Schnittstelle von Technologie und Menschenrechten sowie der Verhinderung von Marginalisierung. Im Büro der OSZE-Beauftragten für Medienfreiheit fokussiert sich Julia auf Policy Research und Development im Bereich Internet-Governance und digitale Partizipation, mit Schwerpunkt auf künstliche Intelligenz und Meinungsfreiheit, digitale und rechtliche Sicherheit von JournalistInnen, sowie Gender und Pluralismus. Zuvor arbeitete sie als Völkerrechts- und Menschenrechtsreferentin im österreichischen Außenministerium. Julia ist Mitglied im Beirat des Wiener Forums für Demokratie und Menschenrechte, ist Magistra der Rechtswissenschaften, hat einen LLM für Informations- und Medienrecht und promoviert über die Auswirkungen digitaler Innovationen auf die Medienfreiheit.

Abstract

The COVID-19 pandemic has been accompanied by reports of unprecedented amount of deceptive and false online information with the potential to severely undermine individual and public health as well as the enjoyment of human rights. Both states and internet intermediaries have undertaken unparalleled steps to address this COVID-19 "infodemic". Indeed, the COVID-19 pandemic may represent a turning point for the governance of the online information landscape in general and the fight against disinformation in particular.

This paper examines responses to disinformation, in particular those involving automated tools, from a human rights perspective. It provides an introduction to current automated content moderation and curation practices, and to the interrelation between the digital information ecosystem and the phenomenon of disinformation. The paper concludes that an unwarranted use of automation to govern speech, in particular highly context-dependent disinformation, is neither in line with states' positive obligation to protect nor with intermediaries' responsibility to respect human rights.

The paper also identifies required procedural and remedial human rights safeguards for content governance, such as transparency, user agency, accountability, and independent oversight. Though essential, such safeguards alone appear insufficient to tackle COVID-19 online disinformation, as highly personalized content and targeted advertising make individuals susceptible to manipulation and deception. Consequently, this paper demonstrates an underlying need to redefine advertising- and surveillance-based business models and to unbundle services provided by a few dominant internet intermediaries to sustainably address online disinformation.

Zusammenfassung

Die COVID-19 Pandemie ist mit Berichten über nie dagewesene Mengen an falschen und irreführenden Online-Informationen einhergegangen, mit dem Potential der Gefährdung der individuellen und öffentlichen Gesundheit, sowie der Wahrung der Menschenrechte. Staaten und Internetintermediäre haben beispiellose Maßnahmen gesetzt, um diese COVID-19 „Infodemie" zu bekämpfen. Tatsächlich kann die COVID-19 Pandemie als Wendepunkt in der Regulierung der Online-Informationslandschaft und insbesondere in der Bekämpfung von Desinformation verstanden werden.

Dieses Paper analysiert die Maßnahmen zur Bekämpfung von Desinformation, insbesondere jene mit Einsatz algorithmischer Systeme, aus menschenrechtlicher Sicht. Es bietet einen Überblick in automatisierte Verfahren zur Moderation und Kuration von Inhalten und in das Wechselspiel zwischen dem digitalen Informationsökosystem und dem Phänomen der Desinformation. Die Arbeit kommt zu dem Schluss, dass ein umfassender Einsatz von Automatisierung zur Inhaltskontrolle – insbesondere von stark kontextabhängigen Inhalten wie Desinformation – nicht im Einklang mit der positiven Verpflichtung von Staaten zum Schutz der Menschenrechte und der Verantwortung von Internetintermediären zur Achtung der Menschenrechte steht.

Das Paper identifiziert erforderliche Schutzmaßnahmen für eine grundrechtskonforme Inhaltsregulierung, wie Transparenz, die Stärkung von Betroffenenrechten, Rechenschaftspflicht und eine unabhängige Kontrolle. Die Arbeit folgert, dass entsprechende Garantien zwar unerlässlich sind, jedoch nicht ausreichen, um das Problem der COVID-19 Desinformation zu lösen. Trotz solcher Schutzmaßnahmen bleiben Individuen aufgrund der hochgradigen Personalisierung von Inhalten und gezielter Werbung anfällig für Manipulation und Täuschung. Das Paper zeigt somit die grundlegende Notwendigkeit auf, werbe- und überwachungsbasierte Geschäftsmodelle zu überdenken und die Dienste einiger marktdominanter Internetintermediäre zu entbündeln, um Online-Desinformation nachhaltig zu bekämpfen.

Table of Contents

A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation[1]

1. Introduction

The advancement of digital technologies has enabled unprecedented opportunities for the realization of human rights, democracy, and sustainable development. Over the course of the last decades, the traditional information ecosystem and corresponding channels of dissemination, such as radio, television and print, have been altered dramatically, affecting the way information is sought, received, and imparted. The expansion of the internet, in particular, has provided unparalleled opportunities for human communication, interaction and the global exchange of ideas and opinions, as well as for access to information. Indeed, the digital progress and social networking technologies have enabled new possibilities to create, disseminate and amplify information at a scale, speed and precision never known before.[2]

Disruptive digital technologies have also, however, exacerbated existing challenges to human rights and brought about new obstacles. The ease and convenience of communication has opened the door for abuse, allowing new methods of control,[3] facilitating the "weaponisation" of information, and enabling disinformation to travel across borders unverified and instantaneously.[4] The global COVID-19 pandemic has accelerated these trends, raising genuine concerns for democratic discourse and free speech online.[5]

---

[1] This paper is based on the master thesis "A Treatment for Viral Deception? Automated Moderation of COVID-19 Disinformation" (under supervisor Prof. Nikolaus Forgó), submitted in October 2021 for the LLM information and media law, University of Vienna (completed in December 2021).

[2] Samantha Bradshaw, Hannah Bailey, Philip N. Howard, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project, Oxford Internet Institute, January 2021; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/02/CyberTroop-Report20-Draft9.%20.pdf, p 13.

[3] Cherilyn Ireton and Julie Posetti, UNESCO, Journalism, 'Fake News' & Disinformation, Handbook for Journalism Education and Training, UNESCO Series on Journalism Education, November 2018, https://unesdoc.unesco.org/ark:/48223/pf0000265552_eng, p 15.

[4] OSCE Representative on Freedom of the Media, Report on the expert roundtable: International law and policy on disinformation in the context of freedom of the media, May 2021; https://www.osce.org/files/f/documents/f/9/488884_1.pdf, p 1.

[5] See, *inter alia*, Human Rights Council, Forty-seventh session, Disinformation and freedom of opinion and expression, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, A/HRC/47/25, April 2021, para 2, and Kalina Bontcheva and Julie Posetti (ed.), Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression, Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet', International Telecommunication Union and UNESCO, September 2020; https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf , p 8.

The pandemic and its related social confinement has forced millions to stay at home, and led to an increasing use of social media,[6] with an estimated increase in internet traffic of up to 70%.[7] With people resorting to the internet to socialize and find reliable information, the pandemic has highlighted the imperative for access to accurate information to make safe decisions – and the potential harm of falsities and deception.[8]

While disinformation is no new phenomenon[9] – it has been blamed for election interference, confusing the public and causing offline violence – the ongoing health crisis illustrates that disinformation affects the health of individuals as well as of democratic societies. By blurring the lines between false and true, disinformation undermines public trust in democratic institutions as well as in independent media, and can fuel societal unrest and violence. Thereby, it disempowers individuals and obstructs their meaningful enjoyment of human rights,[10] flourishing when human rights are constrained, and when media independence, plurality or quality are weak.[11] While disinformation holds the potential for harm at all times, it typically becomes more prevalent and more dangerous at times of crises, be it related to the coronavirus, climate or conflict.[12]

Due to its harm on individuals and societies alike, disinformation requires concrete countermeasures. The COVID-19 "infodemic"[13] significantly increased pressure on governments to find regulatory responses and on internet intermediaries to intervene proactively. Responses by states, however, have been identified as problematic, and intermediaries' measures as inadequate.[14]

This paper examines the responses to COVID-19 online disinformation, in particular those involving automated tools, from a human rights perspective. Following an introduction to the digital information ecosystem, automated content moderation, and the phenomenon of disinformation, the paper provides an overview of applicable

---

[6] European Commission, Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling COVID-19 disinformation – Getting the facts right, JOIN(2020) 8 final, June 2020; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008, p 1.

[7] Mihalis Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty – The ongoing case of the COVID-19 'infodemic' , Disinformation and Digital Media as a Challenge for Democracy, pp 367-387, Intersentia, Vol. 6, June 2020, p 368.

[8] UNESCO, The Right to Information in Times of Crisis: Access to Information – Saving Lives, Building Trust, Bringing Hope!, World Trends in Freedom of Expression and Media Development, June 2020; https://unesdoc.unesco.org/ark:/48223/pf0000374369, p 2.

[9] Different historical sources see disinformation dating back to ancient Rom, where Octavian launched a smear campaign against Antony and Cleopatra, see UNESCO, Journalism, 'Fake News' & Disinformation, p 15 and sources therein. Others even refer to 500 BC and an Athenian naval commander intentionally misleading the Greek troops to join the Persians, see Judit Bayer, Natalija Bitiukova, Petra Bárd, Judit Szakács, Alberto Alemanno, Erik Uszkiewicz, Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States, European Parliament, study requested by the LIBE committee, February 2019; https://www.europarl.europa.eu/RegData//etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf, p 60 and sources therein.

[10] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 2.

[11] *Ibid*, para 4 and para 84.

[12] See, for example, the European Union's information sheets on fighting disinformation in times of crisis, https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation_en.

[13] For a definition of this term, see footnote 267 and 277.

[14] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 3.

international human rights law and assesses their implications as a result of states' and intermediaries' responses to COVID-19 disinformation.[15] While there is evidence that the COVID-19 pandemic could be a turning point for media consumption habits, content moderation practices and the fight against disinformation, the pandemic also highlights the human rights shortcomings of automated moderation of COVID-19 disinformation on large internet intermediaries' services.

2. Digital information ecosystem

2.1. The "platformization" of communication

The internet and online services contribute significantly to a vibrant and robust democratic exchange.[16] Since the advancement of the early internet, the scale of content generated and uploaded by users of online platforms has risen dramatically.[17] This trend has contributed to converge on various fronts. Online content is increasingly blurring the lines between information and entertainment,[18] as well as fact and opinions. News and public interest content, too, are increasingly merged with information lacking the accountability structures that characterize legacy media.[19]

Both private communications and public debate increasingly take place on online platforms with extensive economic, infrastructural and cultural extensions, a phenomenon often referred to as "platformization".[20] The more direct access to content via platforms search engines – without editorial mediation by journalists – has additionally transformed the consumption of information, and possibilities for persuasion and manipulation.[21]

In contrast to the stark decentralization in the production of information, the digital ecosystem itself has enabled a maximum concentration of a few private actors' power to

---

[15] This paper focuses on the international human rights framework and case law by the European Court of Human Rights. It does not assess in detail existing national legislation on content moderation or regulatory proposals such as the European Union Digital Services Act, nor case law by the European Court of Justice.

[16] Already in 2011, the UN Special Rapporteur on freedom of expression underlined the potential of the internet to be a catalyst for individuals to exercise free speech, which, in turn, is an "enabler" for other rights. See Human Rights Council, Seventeenth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 16 May 2011, A/HRC/17/27; https://undocs.org/A/HRC/17/27, para 23.

[17] Carey Shenkman, Dhanaraj Thakur, Emma Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology (CDT), May 2021; https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf, p 10.

[18] Maximilian Gahnz, Katja T. J. Neumann, Philipp C. Otte, Bendix J. Sältz, Kathrin Steinbach, Breaking the News? Politische Öffentlichkeit und die Regulierung von Medienintermediären, Friedrich Ebert Stiftung, February 2021, p 2.

[19] Damian Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, Artificial Intelligence – Intelligent Policies: Challenges and opportunities for media and democracy, Background Paper, Council of Europe, Ministerial Conference, February 2020; https://rm.coe.int/cyprus-2020-new-media/16809a524f, p 5f.

[20] OSCE Representative on Freedom of the Media, Non-paper on the Impact of Artificial Intelligence on Freedom of Expression, March 2020; https://www.osce.org/files/f/documents/b/a/447829.pdf.

[21] Carme Colomina, Héctor Sánchez Margalef, Richard Youngs, The impact of disinformation on democratic processes and human rights in the world, European Parliament, study requested by the DROI subcommittee, April 2021; https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf, p 2.

govern the abundance of information online.[22] These gatekeepers typically determine which content is allowed, how it is organized and presented to which audiences, and at what point in time. Their design shapes what is possible online, their policies influence what is permissible, and their personalization algorithms determine what is visible.[23]

Intermediaries' content policies inevitably reflect legal, social and corporate values,[24] and tradeoffs between jurisdictions and scalability.[25] It is thus relevant who has influence over the design, deployment and enforcement of such law-like policies, and how decisions are made. It is significant if these policies that are applied across regions and value systems are predominantly determined by U.S.-based enterprises and white businessmen.[26]

Moreover, in today's digital ecosystem, the same actors that host content also control how it is distributed, whilst simultaneously offering the surrounding advertising infrastructure. Through the monetization of content and user engagement, advertising business models therefore inform content governance policies. Corporate imperatives for profit and growth, however, may be at odds with the public interest and can result in censorship of legitimate expression of personal views on the one hand, and the spread of false information, hate campaigns, and the sowing of insecurity and fear on the other hand.[27]

2.2. Automated content governance

Within one minute, almost 350,000 Instagram stories and almost 150,000 photos are uploaded onto Facebook.[28] Each second, almost 10,000 tweets are posted, 95,000 Google searches made, and 90,000 YouTube videos watched.[29] The enormous amount of content – including user-generated content – available on internet intermediaries' services necessitates some sort of organization and moderation. It also raises the question of liability for own and third-party content, in particular if content – or its visibility and accessibility – is interfered with.

Current liability models for intermediaries' content governance decisions recognize that liability for third-party content and even the fear of being held legally liable can

---

[22] Antonella Sciortino, Fake News and Infodemia at the Time of Covid-19, Direito Publico, Vol. 17, no. 94, pp 35-49, August 2020; https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/4823, p 45.
[23] Barrie Sander, Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, Fordham International Law Journal, Vol. 43, No. 4, pp 939-1006, May 2020; https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3, in particular p 982.
[24] Hannah Bloch-Wehba, Automation in Moderation, Cornell International Law Journal, Volume 53, Issue 1, pp 42-55, Texas A&M University School of Law, March 2020; https://scholarship.law.tamu.edu/facscholar/1448, p 46.
[25] Chris Marsden, Trisha Meyer, European Parliament, Study on regulating disinformation with artificial intelligence, European Parliamentary Research Service, Scientific Foresight Unit, March 2019; https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf, p 16.
[26] Tarleton Gillespie, Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, January 2018, p 12.
[27] Andrei Richter, Fake News and Freedom of the Media, Journal of International Media & Entertainment Law, Volume 8, Number 1, pp 1-34, 2019; https://www.swlaw.edu/curriculum/law-review-journals/journal-international-media-entertainment-law, p 2.
[28] Domo, Data Never Sleeps 8.0, April 2020; https://www.domo.com/learn/infographic/data-never-sleeps-8.
[29] World Wide Web Consortium (W3C), World Wide Web Foundation, Internet live stats; https://www.internetlivestats.com.

result in precautionary removals of content. In order to avoid the chilling of the free flow of information and exchange of ideas,[30] intermediaries are thus given the right but not the responsibility to moderate content[31] – as long as they act "in good faith"[32] or lack knowledge of illegality – to incentivize the takedown of problematic content upon notice, but not proactively to avoid collateral censorship.[33]

While this approach stimulated innovation and might have enabled today's social networking, it potentially also contributed to increased corporate power [34] and resulted in intermediaries adopting a laissez-faire approach towards content governance.[35] Over the course of the last decade, these liability models have been widely debated and reconsidered.[36] While some novel approaches demand sophisticated filtering measures, others such as the EU's Digital Services Act call for systemic responsibilities rather than liability models for individual content governance decisions.[37]

In this paper, content governance is understood as intermediaries' policies of the use of, and participation in, their platforms and services. This entails policies on the permissibility and evaluation of content in view of detecting and removing problematic content (content moderation in the narrow sense) as well as design decisions determining the organization and visibility of content (content curation).[38]

Content governance inevitably involves complex decisions about speech, making it a difficult and resource-intensive task. As the abundance of information outpaces any ability for human filtering or ranking, intermediaries increasingly rely on automation to help sort and analyze content, as well as to enforce their rules.[39] Today, content governance is largely implemented by automated and algorithmic systems that determine how widely, at what time and with which audiences each specific piece of content is shared, and whether or not it is removed or promoted to the exclusion of other information.[40]

---

[30] In particular Section 230 of the U.S. Communications Decency Act and the EU E-Commerce Directive. See, for example, Fernando Nuñez, Disinformation Legislation and Freedom of Expression, UC Irvine Law Review, Vol. 10, Issue 2, March 2020; https://scholarship.law.uci.edu/ucilr/vol10/iss2/10, p 790.

[31] Gillespie, Custodians of the Internet, p 30ff, p 44.

[32] Bloch, Automation in Moderation, p 49.

[33] European Parliament, Study on regulating disinformation with artificial intelligence, p 24. In Europe, there is a ban on the general monitoring of content as monitoring system obligations would be "capable of undermining the right to impart information on the internet". See European Court of Human Rights, Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, application no. 22947/13, judgment, 2 Feburary 2016; http://hudoc.echr.coe.int/eng?i=001-160314.

[34] Bloch, Automation in Moderation, p 50.

[35] Nicolas P. Suzor, Lawless: the Secret Rules That Govern Our Digital Lives, Cambridge University Press, 2019.

[36] European Parliament, Study on regulating disinformation with artificial intelligence, p 23.

[37] Emma Llansó, Joris van Hoboken, Jaron Harambam, Artificial Intelligence, Content Moderation, and Freedom of Expression, Transatlantic Working Group, February 2020; https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf, p 12f.

[38] Content moderation decisions regarding the permissibility of content are not necessarily binary (i.e., removal of a piece of content or not), but might affect content curation in terms of visibility and prioritization. To avoid the ambiguous differentiation between content moderation and curation, this paper refers to *content governance* as a broad term, encompassing all rules regarding the permissibility, organization and presentation of content.

[39] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 3.

[40] United Nations General Assembly, Seventy-third session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "artificial intelligence", David Kaye, A/73/348, August 2018; https://undocs.org/pdf?symbol=en/A/73/348, p 6.

Automation, and artificial intelligence (AI) in particular, offers speed and scale.[41] It enables a broader and faster sharing and accessing of information,[42] proactive detection and reactive information analysis, as well as the curation of massive amounts of content.[43] While there is no universally agreed definition of AI, it is regularly used as an umbrella term for automated, data-driven processes that exhibit human-like behaviors on a predefined task.[44]

While artisanal content governance (smaller-scale case-by-case content review, e.g., at Medium) and community-reliant content governance (typically combining a small team of human moderators with voluntary community review, e.g., at Reddit) regularly involve little to no automation, large internet intermediaries such as Facebook[45] (including Instagram) and Google (including YouTube) employ industrial content governance, which typically combines automated tools with large-scale human reviewers.[46] Given the lack of transparency about the use of automation, however, it is unclear how much content governance is based on automated decision-making.[47]

As content governance can involve a variety of automated tools and processes at different stages, ranging from simple keyword filters and human-designed instructions to sophisticated machine learning,[48] this paper uses the term automation to capture all automated data-analysis-based processes while focusing on content governance at scale deployed by dominant internet intermediaries. Following descriptions of automated, algorithmic and AI-based tools typically involved in content governance processes, the paper will describe their interconnection with targeted and surveillance-based advertising, followed by a chapter on the shortcomings of automated content governance with specific case studies and illustrating the impact on legacy media.

Industrial content governance increasingly relies on machine learning. As a sub-category of AI, this describes the self-learning training of algorithms to make data-driven predications by progressively identifying new problems and developing new answers,

[41] Daphne Keller, Internet Platforms, Observations on Speech, Danger, and Money, National Security, Technology, and Law, Aegis Series Paper No. 1807, Hoover Institution, June 2018; https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf, p 6.

[42] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 3.

[43] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 2.

[44] Ofcom, Cambridge Consultants, Use of AI in Online Content Moderation, July 2019; https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf, p 16ff. Julia Haas, Freedom of the Media and Artificial Intelligence, Global Conference for Media Freedom, November 2020; https://www.international.gc.ca/campaign-campagne/assets/pdfs/media_freedom-liberte_presse-2020/policy_paper-documents_orientation-ai-ia-en.pdf, p 2.

[45] The Facebook Company renamed itself Meta in late 2021, but the social networking platform retained its name.

[46] Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches, Data & Society, November 2018, https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf, p 6.
More details on artisan on p 17ff, on community-reliant on p 20ff and on industrial on p 23ff.

[47] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 8.

[48] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 3. Council of Europe, Algorithms and Human Rights: Study on the human rights dimension of automated data processing techniques and possible regulatory implications, Council of Europe study DGI(2017)12, prepared by the committee of experts on internet intermediaries (MSI-NET), March 2018; https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5, p 6.

thus constantly adapting itself.[49] It is a process for parsing data to extract features, correlations and relations, without being explicitly programmed to do so, and then applies those understandings to analyze other data.[50] Supervised machine learning is used to classify data based on labeled outputs, involving annotated training data to identify features to categorize input, while unsupervised machine learning aims to understand the structure of datasets in the absence of labels, identifying underlying patterns. Reinforced machine learning builds on feedback and the environment rather than existing data (as, for example, in games to maximize the score), often enabled by deep neural networks that enable recognizing features in complex data inputs ("feed forward" layers describe processes where the output of one layer is simultaneously the input to the next layer).[51] Machine learning is often used in matching tools to recognize content as identical or sufficiently similar to another piece of content and in prediction to recognize the nature of a content item on the grounds of the tool's prior learning.[52] Regardless of its level of sophistication, any contemporary machine-learning tool provides automation only in a specific domain, the computer code remains to be designed by humans, as are the instructions, and the objectives of the application, and regularly the selecting and labeling of input data or output classifications.[53]

[49] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 4 and p 6.
[50] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 12.
[51] Ofcom, Use of AI in Online Content Moderation, p 18ff and glossary p 75f and Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 4. David Leslie, Christopher Burr, Mhairi Aitken, Josh Cowls, Mike Katell, Morgan Briggs, Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, The Alan Turing Institute, March 2021; https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf, p 8.
[52] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 12.
[53] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 4.

Matching technologies provide a tool to identify content against an existing reference database,[54] e.g., to match uploads against existing databases of child sexual abuses, terrorist content or, most recently, COVID-19 disinformation.[55] Matching technologies typically involve hashing, a mathematical function where a content item is uniquely identified to create a digital fingerprint. Cryptographic hashing generates a random hash fingerprint that is sensitive to any change, so that the most minor change would result in a completely different hash value.[56] Conversely, perceptual hashing intends to detect whether content is "alike enough" by computing "homologies" so that similar inputs create similar hash values based on perceptually salient features to capture a distinctive pattern. This makes the hash more resilient and ensures that the content remains identifiable.[57]

Classification or prediction technologies assess content by identifying categories, by the induction of statistical patterns of data and predict outcomes. Such models are typically trained on manually labeled large datasets,[58] inducing generalizations about features based on the examples provided within a given category.[59] Predicting technologies are regularly used to detect linguistic features, attempting to infer context, or used for keyword filtering.[60] In the context of multimedia content analysis, prediction technologies aim to identify characteristics of a content item by generalizing its attributes, e.g. based on classifiers, object detectors, semantic or instance segmentation (relationships between objects), scene understanding or object tracking.[61]

The types and stages of automated tools vary in the content governance process.[62] Automation can be deployed either prior to a content item's upload (pre-moderation), after its publication (post-moderation), or if flagged (reactive moderation).[63]Decision-making processes can be fully automatized (Facebook, for example, declares to

[54] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 14.
[55] *Ibid*, p 42.
[56] Robert Gorwa, Reuben Binns, Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big Data & Society, February 2020; https://doi.org/10.1177/2053951719897945, p 4.
[57] For more details, see Gorwa et al., Algorithmic content moderation, p 4; CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 13 and p 37f (appendix); Ofcom, Use of AI in Online Content Moderation, p 48ff; and Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 6.
[58] Gorwa et al., Algorithmic content moderation, p 5 and CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 19 and more in detail p 43f in appendix.
[59] For more technical details, see Gorwa et al., Algorithmic content moderation, p 5. For an overview of which algorithmic content moderation systems are deployed by major social media platforms, see p 6 and p 7.
[60] Ofcom, Use of AI in Online Content Moderation, p 6.
[61] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 17ff.
[62] For an explanation of the AI lifecycle itself, see The Alan Turing Institute, Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, p 9ff.
[63] For more information on the different stages, see the figure in Ofcom, Use of AI in Online Content Moderation, p 35.

automatically remove content if the prediction level indicates the decision is more accurate than a human review[64]) or quasi- or semi-automated, when automated tools prepare human decisions.[65]Automated tools can be used to flag content for review, either in queue with user-flagged content or prioritized for a quicker review or allocating content to specific human moderators, or to directly remove or block content from its upload.[66] Automated tools can also be deployed to subject users to warnings, suspension or deactivation.

In addition to the use of automation to detect, evaluate and act upon content, it can also be used to optimize content governance processes themselves. To this end, automated tools can be used to synthesize data to improve training and moderation performance, for instance by creating harmful content through 'style transfer' to supplement existing examples. Automation can also be used to generate more data out of a limited set, to repair incomplete data, or to correct bias in datasets, for example by generating additional data of an underrepresented group.[67] Automation can also facilitate the reporting of problematic content,[68] and support human moderators, by prioritizing or varying content, or by limiting exposure or blurring specific areas to reduce harmful effects.[69] Besides, automation can be deployed to encourage socially-positive engagement of individuals. The prevalence of harmful content online is often associated with factors like anonymity, empathy deficits, or asynchronous communication. Hence, automated nudging techniques can be used to suggest alternative language or a short delay before responding, or to highlight potentially harmful content.[70]

2.3. Content governance and advertising business models

Newsfeeds, video recommendations and search engines alike build on automated content recommender systems that determine what is seen and what remains hidden, either following an active user query (as in a search engine) or a personalization independent from explicit user input.[71] Individualized ranking and content selections enable individuals to navigate through the abundance of information online and thus follows the economic necessity of creating a user-friendly environment.[72]

---

[64] Sanders, Human Rights-Based Approach, p 947.
[65] Council of Europe, Algorithms and Human Rights, p 3.
[66] Gorwa et al., Algorithmic content moderation, p 6.
[67] For more information and concrete examples, see Ofcom, Use of AI in Online Content Moderation, p 58ff.
[68] Ofcom, Use of AI in Online Content Moderation, p 67ff.
[69] *Ibid*, p 60f.
[70] Nudging is, however, triggering own debates about ethics and freedom of choice In particular if choice architectures are manipulated to incentivize socially-positive behavior, are not rigorously transparent or without consequences to make alternative decisions.
[71] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 15.
[72] Kate Klonick, The New Governors of Speech: The People, Rules, and Processes Governing Online Speech, Harvard Law Review, Vol. 131, No. 6, pp. 1598-1620, April 2018; https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech, p 1669.

No ranking, however, is "neutral". Just as decisions about content removals, content recommendations inevitably involve value-laden judgments. Recommender systems are designed, developed and deployed in a specific context, by a specific actor, and for a specific purpose, entailing commercial and political considerations.[73] The Council of Europe, for example, recognized that platforms' prioritization of certain values over others shapes contexts in which individuals access and process information, and make conclusions and decisions.[74] Intermediaries recommend content based on "relevance", identified by matching profiles with a large pool of content to compute a "similarity score" between the individual user profile and characteristics of each single piece of content.[75] Consequently, each individual is presented with a different curated segment of the online discourse based on their personal profile.

This personalization leads to a fragmentation of information and "publics", often referred to as filter bubbles or echo chamber.[76] While individuals have to an extent always been exposed to divergent information, the vectors, volumes and velocity have increased significantly.[77] In today's digital ecosystem, each person has access to theoretically infinite information, but in reality, the exposure to information varies substantially. Consequently, divided parallel realities and narratives may emerge, which results in a diminishing amount of agreed facts.[78] The personalized experiences online have been named to trap individuals in "informational cages" that are built on sophisticated profiling techniques and opaqueness.[79] While the individual and democratic impact of the fragmented information landscape is difficult to assess, they have been associated with questionable manipulations of individuals, including during the 2016 U.S. presidential elections, the UK's Brexit referendum, the Cambridge Analytica scandal,[80] and during the COVID-19 pandemic.

Today's content curation builds on the profiling and micro-targeting of individuals (users and non-users alike), which simultaneously serves intermediaries' advertising purposes. While hosting and distributing user-generated content may be the more apparent service, the main revenue of large intermediaries stems from buying, selling

---

[73] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 16.

[74] Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, 1337th meeting of the Ministers' Deputies, Decl(13/02/2019)1, February 2019; https://search.coe.int/cm/pages/result_details.aspx?ObjectId=090000168092dd4b, para 7.

[75] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 14.

[76] Both terms *filter bubble* and *echo chamber* describe a partial information blindness, the first referring to the narrowing of the choice of content visible to users, and the latter referring to limited exposure to diverse content due to the prominence and recommendation of content that reinforces the users' existing views.

[77] European Parliament, Study on regulating disinformation with artificial intelligence, p 12, this has also been recognized for search engines, see Human Rights Council, Thirty-second session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 11 May 2016, A/HRC/32/38; https://undocs.org/en/A/HRC/32/38, para 21.

[78] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 17. At the same time, however, some studies indicate that emotional content ma spark sustained participation reducing information bubbles. See, Sarah Shugars, Nicholas Beauchamps, Why Keep Arguing? Predicting Engagement in Political Conversations Online, SAGE, March 2019; https://doi.org/10.1177/2158244019828850.

[79] Sciortino, Fake News and Infodemia at the Time of Covid-19, p 38. For an illustrative approach to how individuals are trapped in caves, see Vladan Joler, New Extractivism, 2020, https://extractivism.online.

[80] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 16.

and marketing for advertising.[81] While the simultaneous collection, processing and monetization of data may not be permitted for some actors,[82] internet intermediaries' business model precisely builds on this coupling by delivering individuals – depicted as mere consumers – to advertisers. Content governance, in a simplified manner, thus follows the logic of likeability to drive traffic and to engage users, which, in turn, sells advertising and generates profit.[83] In order to maximize profits in an ecosystem where information is abundant, intermediaries thus sell human attention.[84]

Consequently, content governance is designed to promote content that is predicated to entice greater reach and user interaction,[85] prioritizing speed and scale over quality, clicks and time spent over public interest.[86] With extreme content increasing the probability of engagement, intermediaries are incentivized not to over-moderate content that may be "valuable",[87] assessed on calculations to attract or repel user engagement.[88] Sensational and controversial content can attract more engagement, just as racism, misogyny, disinformation or content instilling fear or hatred can.[89] Consequently, while not allowing all highly "clickable" content to avoid reputational damage for too much harmful content, content recommender systems may drive individuals towards emotionally charged, deceptive or inflammatory content, rewarding problematic content for attracting attention by further increasing its reach.[90]

In order to predict the "value" and "relevance" of content, intermediaries depend on fine-grained data. In exchange for providing their services "for free", intermediaries surveil individuals' online behavior, and monetize the harvested and analyzed data through targeted advertising.[91] Although the tracking and profiling for personalized marketing and automated distribution systems for sponsored content differs in nature, they are closely interlinked with the curation of organic user-generated content.[92]

---

[81] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising: The case against commercial surveillance online, June 2021; https://www.forbrukerradet.no/wp-content/uploads/2021/06/20210622-final-report-time-to-ban-surveillance-based-advertising.pdf, p 21.

[82] Doctors, lawyers and investment brokers, for example, also potentially collect and deal with large amounts of sensitive data, but are not permitted to monetize this data in their own interest. See European Parliament, Disinformation and propaganda, p 90.

[83] Sarah T. Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation, First Monday, Vol. 23, No. 3-5, March 2018; https://doi.org/10.5210/fm.v23i3.8283.

[84] Sanders, Human Rights-Based Approach, p 952f. This phenomenon is often referred to as "attention economy".

[85] Arguably, the deployment of automated tools with such optimization goals at scale may result in a self-fulfilling prophecy if content predicted to enlarge user engagement gets amplified. See Ofcom, Use of AI in Online Content Moderation, p 45; and European Parliament, Study on regulating disinformation with artificial intelligence, p 16.

[86] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 17f.

[87] Ofcom, Use of AI in Online Content Moderation, p 62.

[88] Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation.

[89] Haas, Freedom of the Media and Artificial Intelligence, p 3.

[90] Zeynep Tufeki, YouTube, the Great Radicalizer, New York Times Opinion, March 2018; https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. Paul Butcher, COVID-19 as a turning point in the fight against disinformation, Nature Electronics, Volume 4, pp 7-9, January 2021; https://doi.org/10.1038/s41928-020-00532-2, p 8.

[91] Nathalie Maréchal, Targeted Advertising Is Ruining the Internet and Breaking the World, VICE: Motherboard, November 2018, https://www.vice.com/en/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world.

[92] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 14.

The technical design for targeted advertising can optimize content curation and the monitoring and censoring of content. It can involve the same labeling of users and content, the same learning patterns, and the same data or optimization logic.[93] As large internet intermediaries' business models are based on ad revenues, they are eliciting more types and greater volumes of data, from increasingly different sources.[94] New computational means enables the inference of most intimate and fine-grained information about individuals, users and non-users alike, in turn, sorting individuals into categories[95] according to demographics and inferred interests.[96] This enables content governance and advertising based on specific characteristics rather than context.[97]

Moreover, increasing user engagement, i.e., time and attention, does not only result in more advertising but also provides access to even more behavioral data and consequently, improved targeted advertisement and increased profit.[98]Content governance also contributes to the vicious circle of data obsession as automated systems encourage even more large-scale processing and profiling, perpetuating the power of those applying these tools.[99] Sophisticated data profiling techniques with opaque design features (dark patterns) can manipulate individuals into accepting even more tracking,[100] which not only raises serious privacy and human rights concerns but also enables anti-competitive behavior.[101]

At the same time, non-commercial advertisers can leverage this industry to their advantage,[102] and the newly created and facilitated surveillance ecosystem can be exploited by states and malicious actors alike, and generally lowers the costs and thresholds for surveillance.[103]

2.4. Shortcomings of automated content governance

While automation per se is not harmful to human rights, its implementation to human interactions, embedded and deployed in a specific context, may have detrimental

---

[93] Niva Elkin-Koren, Maayan Perel, Separation of Functions for AI: Restraining Speech Regulation by Online Platforms, Lewis & Clark Law Review, Vol. 24, Issue 3, pp. 857-898, August 2020; https://dx.doi.org/10.2139/ssrn.3439261, p 8.
[94] Gillespie, Custodians of the Internet, p 19.
[95] Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, para 6.
[96] Dipayan Ghosh and Ben Scott, Digital Deceit: The Technologies Behind Precision Propaganda on the Internet, New America, January 2018; https://www.newamerica.org/pit/policy-papers/digitaldeceit, p 5.
[97] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising, p 5f.
[98] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 10. Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation, September 2021; https://assets.mofoprod.net/network/documents/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation_T2pk9Wj.pdf.
[99] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 10 and Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation.
[100] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising, p 13.
[101] *Ibid*, p 11.
[102] European Data Protection Board, Opinion 3/2018 on online manipulation and personal data, March 2018; https://edps.europa.eu/sites/default/files/publication/18-03-19_online_manipulation_en.pdf, p 10.
[103] OSCE Representative on Freedom of the Media, SAIFE Policy Manual, December 2021; https://www.osce.org/files/f/documents/8/f/510332_0.pdf, p 97 ff. Bloch, Automation in Moderation, p 79ff.

impacts.[104] Due to the prevalent opaqueness and unaccountability of automated content governance processes, the full extent of their impact on human rights and democratic discourse remains unknown. By increasingly deploying automation, including for reasons of cost-effectiveness and alleged streamlining of decisions, intermediaries eliminate "the very human reflection that leads to pushback, questioning and dissent",[105] evading humans' tacit knowledge that enables them to understand exceptional cases where a rule's application may not be appropriate despite falling within its scope.[106]

Taking humans out of the loop of content governance exacerbates existing procedural challenges, including by further increasing opacity (decreasing decisional transparency and auditability), complicating principles of fairness and justice, and re-obscuring the political nature of speech decisions.[107] If unchecked, automated systems increase the remoteness and unaccountability of decision making.[108] The lack of transparency and explainability is often referred to as the black box phenomenon,[109] as data is fed into an automated box and results are extracted while the reasoning behind these results is concealed.[110]

The biggest limitation of automated tools remains their lack of understanding of context. In almost any content decisions, however, context is decisive while also difficult to assess, particularly as rules are often imprecise or ambiguous.[111] Today's automation has a limited ability to assess the nuanced meanings of human communication, or societal, historical, political and cultural context,[112] let alone the motivation of a speaker.[113]

Given the limited ability to assess context, the global deployment of automated content governance tools may lead to bias at the expense of local adaption,[114] especially if tools are trained on data from one region.[115] Technology, moreover, does not perform equally well across languages, cultures, or groups of society.[116] The performance of speech detection algorithms, overall, is significantly lower in languages other than

---

[104] Council of Europe, Algorithms and Human Rights, p 8.
[105] Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation.
[106] Council of Europe, Algorithms and Human Rights, p 9.
[107] Gorwa et al., Algorithmic content moderation, p 1.
[108] European Data Protection Board, Opinion 3/2018, Opinion on online manipulation and personal data, p 5.
[109] Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information, Cambridge, London, Harvard University Press, January 2015.
[110] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising, p 12.
[111] Human Rights Council, Thirty-eighth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, A/HRC/38/35, April 2018; https://undocs.org/A/HRC/38/35, para 2.
[112] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 7, Ofcom, Use of AI in Online Content Moderation, p 4 and CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 29f.
[113] Natasha Duarte, Emma Llansó, Anna Loup, Mixed Messages? The Limits of Automated Social Media Content Analysis, Center for Democracy & Technology (CDT), November 2017; https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf, p 8.
[114] Ofcom, Use of AI in Online Content Moderation, p 42. Caplan, Content or Context Moderation?, p 25.
[115] OSCE Representative on Freedom of the Media, #SAIFE – Spotlight on Artificial Intelligence and Freedom of Expression, July 2020; https://www.osce.org/files/f/documents/9/f/456319_0.pdf.
[116] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 8.

English.[117] While an automated disregard for context impedes human rights, also a geographic disparity in the application and enforcement of policies, be it based on provided resources and investment, or deliberate inaction in complex settings, may lead to other serious infringements.[118]

Besides risks stemming from indiscriminate or discriminatory application of automated content governance tools,[119] biases are regularly be built into them.[120] Biases can be intentional or unconscious, and can stem from bad data quality (encoded biases, prejudiced assumptions or insufficient data)[121] or specific interests of the organizations behind them.[122] Biases are inherent in non-representative (over- or under-inclusive[123]) datasets or can arise in the process of labelling training datasets,[124] or a shortage of suitably qualified staff.[125] All of this risks perpetuating or exacerbating indirect discrimination through stereotyping[126] and reinforcing social bias,[127] with marginalizing effects on minority populations.[128]

Used at scale and with growing pervasiveness, automated systems can encapsulate prejudices and biases, directly impacting individuals' rights,[129] including due to biases regarding protected categories such as race, gender or age,[130] with intersecting layers that are often not accounted for in anti-discrimination frameworks.[131]

While some limitations might be addressed by future technological advances, some are inherent to automation, especially if applied at scale and in the absence of

---

[117] Raxona Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, Social Media + Society, Volume 6, Issue 3, July-September 2020; https://doi.org/10.1177%2F2056305120948190, p 3.

[118] For an example of geographic disparities and their implications, see BuzzFeed News, Craig Silvermann, Ryan Mac, Pranav Dixit, "I Have Blood on My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation September 2020; https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo.

[119] UN Special Rapporteur on freedom of expression, A/HRC/38/35, para 10.

[120] For an early attempt to identify ways to promote equality and non-discrimination in machine learning, see Toronto Declaration, May 2018; https://www.torontodeclaration.org.

[121] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, para 8f. Council of Europe, Committee of experts on internet intermediaries (MSI-NET), Study on the Human Rights Dimension of Automated Data Processing Techniques (in particular Algorithms) and Possible Regulatory Implications, MSI-NET(2016)06 rev3 final, October 2017; https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a.

[122] Eleonora Maria Mazzoli and Damian Tambini, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, Council of Europe, DGI(2020)19, November 2020, https://rm.coe.int/publication-content-prioritisation-report/1680a07a57, p 29.

[123] Gorwa et al., Algorithmic content moderation.

[124] Ofcom, Use of AI in Online Content Moderation, p 41

[125] *Ibid*, p 25ff.

[126] Council of Europe, Algorithms and Human Rights, p 28.

[127] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 4. Picture analysis, for example to detect nudity, are vulnerable to misclassification of underrepresented skin tones, see p 6.

[128] CDT, Mixed Messages? The Limits of Automated Social Media Content Analysis, p 15.

[129] Yifat Nahmias, Maayan Perel, The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, Harvard Journal on Legislation, Vol. 58, no. 1, pp. 146-194, February 2021; https://harvardjol.com/wp-content/uploads/sites/17/2021/02/105_Nahmias.pdf, p 147f.

[130] Council of Europe, Algorithms and Human Rights, p 26.

[131] Julia Angwin and Hannes Grassegger, Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children, ProPublica, June 2017; https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

accountability.[132] All externalities are also closely linked to the power concentrated with a few intermediaries,[133] while specific challenges can arise in the context of smaller intermediaries with fewer human and financial resources,[134] or less expertise, access to datasets and computational power.[135]

In any case, the widespread overconfidence in technology for solving societal and political challenges entrenches human rights risks and existing power dynamics,[136] potentially fueled by intermediaries overselling technological possibilities to avoid regulation.[137] Even if following a "logic of opacity", automated decision-making is regularly perceived as more neutral and objective than that of humans.[138] Such a "tech solutionism" disregards substantial concerns about algorithmic control, surveillance, collateral censorship, and private power.[139] It disregards the current digital ecosystems' power to replicate and amplify many of capitalism's most problematic settings.[140]

2.5. Specific case studies of automated analysis of text, images and multimedia

Automated content moderation tools are widely used to detect and act upon illegal and otherwise harmful content, either to decide about the permissibility of a certain piece of content ex ante, for example by hash matching for image recognition prior to its upload (e.g., PhotoDNA or Content ID), or ex post, for example through a language evaluation following community flagging.[141] While ex ante tools may be described as privatized digital prior restraint,[142] most governance tools employ a "publish-then-filter" approach.[143]

Automated tools are generally well-suited for analyzing known and existing content[144] such as proactively screening content against a database of terrorist content against a database. The Christchurch attack in March 2019, which was accompanied by 1.5 million video uploads within 24 hours, constituted a turning point in this regard. The

---

[132] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 35f.
[133] Klonick, The New Governors of Speech: The People, Rules, and Processes Governing Online Speech.
[134] Gillespie, Custodians of the Internet, p 71.
[135] Ofcom, Use of AI in Online Content Moderation, p 64f.
[136] Bloch, Automation in Moderation, p 81f.
[137] This is closely interlinked with capabilities to lobby hard to shape regulation, by far the EU's biggest lobby spending industry, see Corporate Europe Observatory, The lobby network: Big Tech's web of influence in the EU, August 2021; https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu.
[138] Council of Europe, Algorithms and Human Rights, p 7 and Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation.
[139] Bloch, Automation in Moderation, p 75.
[140] Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation.
[141] For an overview of different filtering techniques, see Giovanni Sartor, Andrea Loreggia, European Parliament, The impact of algorithms for online content filtering or moderation: "Upload filters", study requested by the JURI committee, September 2020;
https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf.
[142] Emma Llansó, No amount of "AI" in content moderation will solve filtering's prior-restraint problem, Big Data & Society, Volume 7, Issue 1, April 2020; https://doi.org/10.1177%2F2053951720920686.
[143] Gillespie, Custodians of the Internet, p 75. To the contrary, app submissions, for example are checked prior to their upload onto Apple and iPhone app stores.
[144] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 16.

Global Internet Forum to Counter Terrorism (GIFCT) holds a hash database of terrorist content to check new uploads against illicit content.[145]
According to YouTube, 98% of terrorist content is automatically prevented from its upload. And while Facebook claims to always provide human review, its transparency reports refer to automatically blocked content.[146] In fact, Facebook refers to 99.6% of terrorist propaganda being proactively removed. This data, however, does not disclose what is considered terrorist or who labels what kind of data, let alone the technical details involved.[147]Moreover, the official numbers of automated removals remain unaudited claims by the intermediaries themselves.[148]

Automated tools struggle, however to consider social, historical, linguistic and other relevant contexts, let alone balance context sensitivities with consistency.[149] Automated tools coded to employ binary decisions, such as whether to remove content or not, may thus increase the number of errors,[150] which is illustrated by the deletion of a Pulitzer-prize winning photograph,[151] the removal of civil war documentation,[152] the facilitation of online incitements linked to mass atrocities,[153] as well as educational material and journalistic content.[154]

At the same time, however, there is increasing government pressure for filtering technologies, namely through non-binding agreements such as the Christchurch Call, the European Commission Recommendation to fight illegal content online, or through regulatory frameworks such as the EU Terrorist Directive (2017/541).

Automated tools are also regularly used to detect copyright-infringing material. YouTube's Content ID, for example, checks new uploads against a database of hashed

[145] Gorwa et al., Algorithmic content moderation, p 2.
[146] Gorwa et al., Algorithmic content moderation, p 6f. Facebook states that it uses human fact-checkers and does not automatically remove false content, but uses automation to support fact-checkers and to identify similar pieces of previously debunked content to take them down quickly and at scale. See Facebook COVID-19 policies, in particular https://about.fb.com/news/2020/03/combating-covid-19-misinformation, https://www.facebook.com/help/230764881494641 and https://about.fb.com/news/tag/covid-19.
[147] Gorwa et al., Algorithmic content moderation, p 12.
[148] European Parliament, Study on regulating disinformation with artificial intelligence, p 17.
[149] Caplan, Content or Context Moderation?, p 13.
[150] Error rates also arise when human moderators decide within a matter of seconds, not allowing complex contexts to be taken into account. See Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation. It is necessary, moreover, to also acknowledge challenges faced by human moderators as content moderation and deciding on horrific content can result in serious trauma.
[151] This refers to the well-known example of the removal of the Vietnam picture of Kim Phuc which was taken down because of child nudity. It is was removed by human reviewer, not an automated system and the removal did not constitute an error *per se* as Facebook's rules did not provide for exemptions to its nudity policy. See, for example, A/HRC/73/348 para 37.
[152] Human Rights Watch, "Video Unavailable": Social Media Platforms Remove Evidence of War Crimes, September 2020; https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.
[153] Human Rights Council, Thirty-ninth session, Report of the independent international fact-finding mission on Myanmar, September 2018; https://undocs.org/en/A/HRC/39/64 and Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar, A/HRC/39/CRP.2, September 2018; https://undocs.org/A/HRC/39/CRP.2.
[154] Bloch, Automation in Moderation, p 65.

copyright material.[155] But again, due to flaws of automated matching to assess context,[156] such detection tools regularly fail to identify ”fair uses”, satire, or other protected speech.[157] Moreover, the alleged infringer who is subject to automated moderation often has little to no recourse to question potentially erroneous decisions,[158] shifting the costs of protecting freedom of expression to individuals. Article 17 of the EU Copyright Directive (2019/790), which de facto requires proactive monitoring of content, and more so, to deploy automated tools like upload filters, has thus been heavily contested.[159]

Compared to the detection of copyright or pornographic content, judgments about what constitutes hate speech or disinformation are even more challenging. Automated tools intended to identify “toxic speech” make predictions about the impact of a specific piece of content. Yet, failing to account for context risks triggering over- and under-zealous predictions.[160] Analyzing newer and more complex formats, such as memes or videos, brings even more limitations to identifying problematic content.[161]

Furthermore, automatic text analysis such as natural language processing tools have been identified to reproduce or even amplify inequalities as well as disproportionally censor already marginalized groups.[162] The aim of eradicating homophobic and transphobic speech, for example, resulted in LGBTQI+ users being censored for their counter-speech or for reclaiming terms.[163]

Keyword filters meanwhile are inevitably overbroad or under-inclusive. They can be easily circumvented,[164] for example by using the code “CV” instead of “COVID-19” to avoid blocking of hateful conspiracy videos.[165] Developers of more sophisticated natural language processing have repeatedly identified shortcomings. The team behind Google/Jigsaw’s Perspective API, which is an open-source toolkit to evaluate the “toxicity” of speech, for example, cautioned against its use due to its errors,[166] inherent

[155] In case of a match, the copyright holder can chose to monitor, block or monetize the upload. See Gorwa et al., Algorithmic content moderation, p 6.
[156] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 15.
[157] Gorwa et al., Algorithmic content moderation, p 7f.
[158] *Ibid*, p 6.
[159] Bloch, Automation in Moderation, p 67f.
[160] Gorwa et al., Algorithmic content moderation, p 9f. European Parliament, Study on regulating disinformation with artificial intelligence, p 6. Identifying “equivalent” content following the identification of problematic speech is even more context-dependent and thus difficult. Against this background, the recent judgment allowing to request the global blocking of identical and “equivalent” defamatory statements may be informed by an overestimate of technology. For the judgment, see European Court of Justice, Eva Glawischnig-Piesczek v. Facebook Ireland Limited, C-18/18, 3 October 2019; https://curia.europa.eu/juris/liste.jsf?num=C-18/18.
[161] For more details on the variety of content formats and associated challenges, including highly complex and contextual content such as memes, see Ofcom, Use of AI in Online Content Moderation, p 32ff.
[162] CDT, Mixed Messages? The Limits of Automated Social Media Content Analysis, p 12ff. At the same time, AI has been used to help identifying bullying material online. For an overview of the image, text and metadata analysis, see Ofcom, Use of AI in Online Content Moderation, p 56f.
[163] Jillian C. York, Corynne McSherry, Content Moderation is Broken, Let Us Count the Ways, Electronic Frontier Foundation (EFF), April 2019; https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways, p 3f.
[164] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 7.
[165] Bellingcat, How Coronavirus Scammers Hide On Facebook And YouTube, March 2020; https://www.bellingcat.com/news/rest-of-world/2020/03/19/how-coronavirus-scammers-hide-on-facebook-and-youtube.
[166] Github, Conversation AI, https://conversationai.github.io.

biases and misclassification disproportionately affecting certain societal groups.[167] In the same vein, an OpenAI research team developing the predictive text tool GPT-2 to generate text released a less capable model due to its risk of producing errors. The newer natural language generation model GPT-3 will only be provided for approved cases[168] as it has led to concerns of misuse, including for automated disinformation at scale.[169]Error rates (both false positives and false negatives) involve serious free speech concerns. False negatives[170] put a significant burden on individuals' rights and may chill speech, lead to self-censorship and silence marginalized voices, while false positives[171] risk limiting legitimate speech.[172] It also needs to be considered whether rendering illegal or unpleasant speech largely invisible may involve other challenges in successfully addressing them.[173]

Using language analysis tools to automatically detect false information may even be more difficult than toxic speech due to the difficulty to parse complex and potentially conflicting meanings of text.[174] Identifying malicious users (and, for example, prioritizing their content for review)[175] or identifying bots, instead, is less error-prone.[176]

With news consumption having increased drastically during the global COVID-19 pandemic, the use of social media, search engines, and other digital media increased sharply, as did public conversation and debates around the pandemic (an analysis of U.S. Twitter demography and activity during 2020 found that almost 10% of all tweets were related to COVID-19[177]). Studies have also shown that COVID-19 online information has included a large number of questionable sources, as well as false information and conspiracy theories.[178] The wish to address challenges associated with the pandemic has triggered a wide use of a variety of automated decision-making systems, such as tracking tools or new content governance tools, based on the idea that every social problem can

---

[167] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 5.
[168] Openai, OpenAI API, June 2020; https://openai.com/blog/openai-api.
[169] CSET, Ben Buchanan, Andrew Lohn, Micah Musser and Katerina Sedova, Truth, Lies, and Automation: How Language Models Could Change Disinformation, May 2021; https://cset.georgetown.edu/publication/truth-lies-and-automation.
[170] The term *false negative* describes an automated misclassification of content that should have been classified as impermissible according to the rules implemented by the automated tool.
[171] The term *false positive* describes an automated wrong assessment of content as objectionable, while in fact it is permissible.
[172] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 9. Haas, Freedom of the Media and Artificial Intelligence, p 3.
[173] Gorwa et al., Algorithmic content moderation, p 11.
[174] European Parliament, Study on regulating disinformation with artificial intelligence, p 2.
[175] Ofcom, Use of AI in Online Content Moderation, p 51 and p 53f.
[176] European Parliament, Disinformation and propaganda, p 33.
[177] For more information on Twitter publics and who tweeted what (e.g. Democrats, African-Americans, younger people etc. overrepresented) and the distribution of engagement, see Sarah Shugars, Adina Gitomer, Stefan McCabe, Ryan J. Gallagher, Kenneth Joseph, Nir Grinberg, Larissa Doroshenko, Brooke Foucault Welles, David Lazar, Pandemics, Protests, and Publics: Demographic Activity and Engagement on Twitter in 2020, Journal of Quantitative Description: Digital Media 1, pp 1-68, April 2021; https://doi.org/10.51685/jqd.2021.002.
[178] Bianca Residorf, Grant Blank, Johannes Bauer, Shelia Cotten, Craig Robertson, Megan Knittel, Information Seeking Patterns and COVID-19 in the United States, Journal of Quantitative Description: Digital Media, Volume 1, April 2021; https://doi.org/10.51685/jqd.2021.003, p 5 and sources therein.

be "fixed" through technology.[179] While it is still too early to fully assess the impact of the use of automation to address the COVID-19 pandemic, this paper aims to assess the human rights implications of automated moderation of COVID-19 disinformation.[180]

2.6. Impact on legacy media

In today's digital ecosystem, legacy media increasingly loses its traditional role as gatekeeper while others bypass the regulatory obligations to which journalists and media outlets are held accountable.[181] Also media's watchdog role as a corrective between powerful actors, such as the government and the public, has been disrupted by the facilitation of direct distribution of information to the public.[182] Whereas legacy media is highly regulated and builds on professional ethics, intermediaries – who disseminate and monetize content generated by others – may curate content primarily by popularity,[183] rewarding engagement rather than diversity or truth. This may result in the prioritization of "click-worthy" information outside established frameworks for journalistic ethics and news accountability.[184]

The structural changes in the way news and media content is gathered and distributed have expanded problems of trust and threaten the historical understanding of media independence as a principle and media responsibility as an expectation.[185] The absence of trust, however, significantly impairs the function of journalism.[186]

In addition, advertising-focused and data-driven automated content governance raises serious media pluralism concerns. The dominance of a few intermediaries simultaneously acting as speech moderators and advertisers additionally pressures legacy media.[187] In short, data-driven advertising is an inherent part of the challenges legacy media currently faces.[188]

The COVID-19 pandemic and associated economic fallout only accelerated the dramatic pressure facing independent journalism, at a time when trustworthy, well-

[179] For more information on this general trend, see, for example, AlgorithmWatch, Automating Society Report 2020, Life in the automated society: How automated decision-making systems became mainstream, and what to do about it, September 2020; https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf.
[180] For responses to COVID-19 online disinformation, see chapter V-VII.
[181] UNESCO, Journalism, 'Fake News' & Disinformation, p 35f.
[182] Richter, Fake News and Freedom of the Media, p 2.
[183] UNESCO, Journalism, 'Fake News' & Disinformation, p 37.
[184] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 5f. Haas, Freedom of the Media and Artificial Intelligence, p 2.
[185] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 5.
[186] *Ibid*, p 7.
[187] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 11.
[188] European Parliament, Study on regulating disinformation with artificial intelligence, p 10.

investigated and fact-checked information is needed more than ever.[189] In the absence of accuracy and facts, however, disinformation fills the void.[190]

3. Concept of disinformation

"A lie can travel halfway around the world while the truth is still putting on its shoes." This quote, ironically, has been falsely attributed to Thomas Jefferson, Winston Churchill, and Mark Twain.[191]

3.1. Marketplace of ideas or just marketplace

Liberal democracies build on the metaphoric concept of the "marketplace of ideas",[192] an ideal of constant dialogue and naturally clashing opinions, where truth ultimately prevails.[193] Any marketplace, however, can fail or be manipulated by power and control.[194]

New technologies provide new possibilities to weaponize information at scale, and to flood the marketplace of ideas with falsehood.[195] Indeed, any disruptive technology – be it the printing press, the radio, television broadcasting, or the internet – can facilitate the amplification of propaganda and hoaxes.[196]

Moreover, today's internet intermediaries face limited quality control standards compared to traditional channels of information distributors such as legacy media.[197] On the contrary, platformized[198] content curation optimized for advertising may even incentivize the promotion of sensational and controversial content to keep users engaged.[199] Filters, either explicit through hashtags, for example, or implicit through automated recommender systems, can be instrumentalized for manipulation.[200] Also

---

[189] Judit Bayer, Bernd Holznagel, Katarzyna Lubianiec, Adela Pintea, Josephine B. Schmitt, Judit Szakacs, Erik Uszkiewicz, Disinformation and propaganda: impact on the functioning of the rule of law in the EU and its Member States – 2021 update, European Parliament, study requested by the INGE committee, April 2021; https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653633/EXPO_STU(2021)653633_EN.pdf, p 67. For an analysis of the impact of the pandemic on journalism, see International Center for Journalists (ICFJ) and Tow Center for Digital Journalism at Columbia University, Journalism and the Pandemic, October 2020; https://www.icfj.org/our-work/journalism-and-pandemic-survey and Reuters Institute, University of Oxford Digital News Report, March 2021; https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.

[190] UNESCO, Journalism, press freedom and COVID-19, World Trends in Freedom of Expression and Media Development, May 2020; https://unesdoc.unesco.org/ark:/48223/pf0000373573?posInSet=1&queryId=0216815c-9a38-457c-8e20-b224c31b03e5, p 15.

[191] Wikiquote, Misquotations, https://en.wikiquote.org/wiki/Misquotations.

[192] A concept first introduced in a dissenting judgment in Abrahams v United States, Case 250 U.S. 616 at 630 (1919), 10 December 1919; https://supreme.justia.com/cases/federal/us/250/616.

[193] Nuñez, Disinformation Legislation and Freedom of Expression, p 788.

[194] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 11.

[195] European Parliament, Study on regulating disinformation with artificial intelligence, p 1.

[196] UNESCO, Journalism, 'Fake News' & Disinformation, p 16.

[197] UNESCO, Journalism, 'Fake News' & Disinformation, p 17.

[198] For a description of the "platformization" of communication, see chapter 2.1.

[199] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 16, see also Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation.

[200] Syed, Real Talk About Fake News: Towards a Better Theory for Platform Governance, pp 345ff.

networked communities built around affinity or ideology[201] and other dynamics such as speed imperatives and bots aiding in maximum persuasion[202] make online discourse prone to manipulation.[203] In short, the sociotechnical context may drive disinformation.[204]

Today's information marketplace resembles a commercial shopping center than a public market agora. Consequently, today's information landscape characterized by algorithmically determined content dissemination may limit vibrant, pluralistic public discourse. This may shift the perception of a free marketplace of ideas – and thus, the emergence of truth.[205]

3.2. Terminology

It is difficult, if not impossible, to identify the clear line between truth and falsity, between legitimate expression aimed at persuading someone and illegitimate manipulation.[206] To date, there is no universally accepted definition of disinformation.[207] Generally, disinformation is referred to as "verifiably false or misleading information that, cumulatively, is created, presented and disseminated for economic gain or to purposefully deceive the public and that may cause public harm".[208] Disinformation thus refers to the spread of falsehood with the intent to deceive and cause harm. It is often orchestrated in an attempt to confuse or manipulate,[209] to harm individual reputations or to incite to discrimination or even violence.[210] Thereby, disinformation interferes with the public's right to know and individuals' right to seek, receive and impart information and ideas of all kinds.[211]

Misinformation, on the other hand, is disseminated unknowingly with no harm meant, while malinformation refers to genuine information presented in an intentionally misleading manner.[212] The impact on society can be similar to the intentional spread of falsehood. While "fake news" is regularly used as catch-all notion encompassing all of

---

[201] *Ibid*, pp 347f.
[202] *Ibid*, pp 350ff.
[203] *Ibid*, pp 348ff.
[204] European Parliament, Study on regulating disinformation with artificial intelligence, p 1. Nabiha Syed, Real Talk About Fake News: Towards a Better Theory for Platform Governance, The Yale Law Journal, Volume 127, pp 337-357, October 2017; https://www.yalelawjournal.org/forum/real-talk-about-fake-news, pp 352ff.
[205] Lombardi, The Illusion of a "Marketplace of Ideas" and the Right to Truth.
[206] Council of Europe, Parliamentary Assembly, Resolution 2143 on "Online media and journalism: challenges and accountability", January 2017; https://pace.coe.int/en/files/23237.
[207] UN SR, HRC Disinformation report para 9f.
[208] The Code of Practice's definition is based on the Report of the Independent High level Expert Group on Fake News and Online Disinformation, See A multi-dimensional approach to disinformation, Report for the European Commission, March 2018; https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation.
[209] UNESCO, Journalism, 'Fake News' & Disinformation, p 7.
[210] Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, March 2017; https://www.osce.org/files/f/documents/6/8/302796.pdf, pp 5.
[211] Joint Declaration on Freedom of Expression and "Fake News", pp 4.
[212] Claire Wardle, Hossein Derakhshan, Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe report, DGI(2017)09, September 2017; https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c, p 20.

the above forms of falsities and deception,[213] it is often instrumentalized to discredit journalists and undermine public trust.[214] "News" in itself refers to "verifiable information in the public interest", so "fake news" can be understood as oxymoron, often used with the intention to weaken the credibility of media content by claiming there is no truth, reliability or impartiality in the news media.[215]

The term deception refers not only to misleading, false, manipulated or fabricated content, but also manipulative actors and deceptive behavior.[216] Propaganda, further, may not even involve false or fabricated information, but rather refers to the use of unethical persuasion techniques for political gain, with the intent to cause insecurity, tear cohesion or incite hostility, or disrupt democratic processes.[217]

In an attempt to capture both deceptive content and behavior, the European Democracy Action Plan refers to "information influence operations", Twitter refers to "platform manipulation", Google to "coordinated influence operation campaigns" and Facebook to "coordinated inauthentic behavior" to also encompass fake engagement by employing bots to increase likes, be it for vanity or deceptive purposes.

These diverging and sometimes even contradicting terms, including in national definitions, may further contribute to lack of conceptual clarity, which risks hampering coordinated international responses.[218] While acknowledging that different terms or concepts are used by states and internet intermediaries, this paper generally refers to disinformation to capture any intentional spread of falsehoods.

3.3. Online disinformation methods

Disinformation is highly contingent on local contexts, power dynamics and cultural environments.[219] It can be shared by individuals or be coordinated by powerful actors.[220] Much like through noise, disinformation can be used to drown out dissenting or critical

[212] Richter, Fake News and Freedom of the Media, p 6.
[214] Independent High level Expert Group on Fake News and Online Disinformation, p 10. See also Joint Declaration on Freedom of Expression and Elections in the Digital Age, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, April 2020; https://www.osce.org/files/f/documents/9/8/451150_0.pdf.
[215] UNESCO, Journalism, 'Fake News' & Disinformation, p 7. Joint Declaration on Freedom of Expression and "Fake News", pp 6; and Richter, Fake News and Freedom of the Media, p 10.
[216] François, ABC.
[217] European Parliament, Disinformation and propaganda, p 18.
[218] UN SR, HRC Disinformation report para 14.
[219] Caplan, Content or Context Moderation?, p 13 and p 25.
[220] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 6 and 30. Reuters, J. Scott Brennen, Felix Simon, Philip N. Howard, Rasmus Kleis Nielsen, Types, sources, and claims of COVID-19 misinformation, April 2020; https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

voices as a "weapon of mass distraction".[221] It can thus be part of influence operations and sophisticated micro-targeted campaigns, spreading beyond borders.[222]

Creators of disinformation regularly make use of psychological means like harnessing emotional responses as anger or disgust,[223] or exploiting existing "data voids".[224] Apart from the actor fabricating information, disinformation can also involve other actors as a separate messenger or distributor.[225] Studies show, moreover, that an increasingly used tactic is to merge falsehood with genuine content in order to increase its reach, especially in private formats, such as messaging apps, social media groups or audio-based messaging apps.[226] Through this technique, disinformation can appear or become organic and create a narrative or even a conspiracy movement.[227] Deceptive behavior also often includes manufacturing virality (using bots, cyborgs or fake accounts[228]) to artificially increase the reach and popularity of certain content for a greater perceived impact.[229] Audiences can be misled by mimicking organic engagement or masking sponsors of messages ("astroturfing") and giving the impression of spontaneous action or support by grassroots participants.[230]

Future technical advances hold the potential to bring new possibilities for disinformation, such as more convincing deepfakes,[231] pervasive virtual reality, or the instrumentalization of virtual assistance and voice-activation.[232] The better automated tools' capacity to predict behavior and influence thoughts will be, the more precise and sophisticated disinformation can be targeted and thus, the greater its risk to manipulate

---

[221] Christina Nemr and William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age, Park Advisors, March 2019; https://www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf; OSCE Representative on Freedom of the Media, Policy Brief on Artificial Intelligence and Disinformation as a Multilateral Policy Challenge, December 2021; https://www.osce.org/files/f/documents/d/0/506702.pdf.

[222] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 6f. External disinformation has so far rather been understood as geopolitical challenge, less as a human rights problem.

[223] Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework.

[224] "Data voids" refers to missing information, for example if search engine queries turn up little to no results. Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 15f.

[225] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression and UNESCO, Journalism, 'Fake News' & Disinformation, p 49ff.

[226] European Parliament, Disinformation and propaganda: 2021 update, p 24.

[227] Camille François, Brookings Podcast on COVID-19 and the ABCs of disinformation, https://www.brookings.edu/techstream/podcast-camille-francois-on-covid-19-and-the-abcs-of-disinformation.

[228] Kate Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, International Law Programme, Chatham House, November 2019; https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf.

[229] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 7. Camille François, Actors, Behaviors, Content: A Disinformation ABC, Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses, Transatlantic Working Group on Content Moderation Online and Freedom of Expression, September 2019; https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf, p 4.

[230] European Parliament, Disinformation and propaganda, p 30. In general, the misuse of intermediaries' services has only recently been recognized as subtle form of attempting to manipulate civic discourse and deceive people instead of as merely from a cybersecurity perspective such as hacking, see François, Actors, Behaviors, Content: A Disinformation ABC, p 2.

[231] Deepfakes are based on real footage to portray a fabricated statement or action by creating detailed mathematical maps of features (encoding) and turning them into new images (decoding), and can lead to a distorted picture of reality. See Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 7.

[232] Future prospects analysis at European Parliament, Disinformation and propaganda, p 116ff.

economic or social choices and behaviors. The peril is particularly high if the targeting takes place subliminally, without knowledge or consent.[233]

At the same time, however, technological advancement that facilitates the production and dissemination of falsehood may also help detect and correct disinformation.[234]

Studies show that already today false information travels online six times faster than true stories, and travels farther and deeper.[235] Although research on the radicalizing effects of content recommender systems remains inconclusive, automated content curation may contribute to this slippery slope by exposing individuals to more and more extreme and deceptive content as they incite stronger feelings and thus engagement.[236]

At the same time, disinformation can be extremely lucrative for their creators and disseminators, for example through monetizing misleading content,[237] as well as for intermediaries providing the advertising and automated targeting infrastructure. Disinformation thus follows the structural and economic logic of online platforms linked to advertising and content curation.[238]

As discussed in the previous chapter, targeting incentivize extensive harvesting and exploitation of data. It also increases the risk of manipulation, as recognized by the UN Special Rapporteur on freedom of expression.[239] In a system where online behavior is constantly commercialized by collecting and exploiting personal information, everyone is vulnerable to manipulation by default.[240] A constant surveillance enables the identification of the most susceptible moment, be it for advertising diet products at times of low self-confidence, marketing gambling when someone struggles with addiction,[241] or providing disinformation to an unsettled person. Targeted, surveillance-based advertising systems may thus contribute to the amplification of disinformation.

Another challenge in the context of targeted advertising is that advertisers regularly do not know where their ads are displayed.[242] Various brands, even if unintentionally, fund conspiracies, including on COVID-19. These brands include Amazon

---

[233] Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic, para 8.

[234] European Parliament, Disinformation and propaganda, p 122.

[235] According to a MIT study analyzing Twitter stories between 2006-2017, false stories are 70 % more likely to be retweeted than true ones. See Soroush Vosoughi, Deb Roy, Sinan Aral, The spread of true and false news online, Science, Vol. 359, Issue 6380, pp. 1146-1151, March 2018, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf.

[236] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 66.

[237] UK House of Commons, Digital, Culture, Media and Sport Committee, Misinformation in the COVID-19 Infodemic, Secord Report of Session 2019-21, July 2020, United Kingdom of Great Britain and Northern Ireland, Parliament; https://committees.parliament.uk/publications/1954/documents/19089/default. Institute for Strategic Dialogue (ISD), Covid-19 Disinformation Briefing No. 1, March 2020; https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-1, p 11f.

[238] European Parliament, Disinformation and propaganda, p 31 and p 74. Ghosh, Digital Deceit, p 29.

[239] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 9.

[240] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising, p 4.

[241] *Ibid*, p 15.

[242] *Ibid*, p 19f.

Prime, Lidl, and even UNICEF.[243] The Global Disinformation Index estimates that in 2020, advertisers provided 25 million USD to nearly 500 coronavirus disinformation websites (excluding social media and video platforms, and in English alone).[244] The top three companies generating ad revenues from such disinformation pages are Google, OpenX and Amazon.[245]

Arguably, intermediaries are the main profiteers from this advertising complexities. Moreover, intermediaries regularly provide infrastructural support to controversial websites, including on COVID-19, for example by providing tracker systems, behavioral analytics, or cross-platform integration tools connected to their services.[246] If false information is linked to the advertising of Google and Facebook, for example, someone visiting a disinformation website may be retargeted with the false content when browsing YouTube or Instagram at a later time.[247] Targeting systems build on trackers collecting information through browser cookies, fingerprinting or IP tracking, and so-called widgets that may carry information across websites.[248]

Google and Facebook are prevalent in all categories of digital advertising, widgets, and analytic trackers, and offer multiple levels of services, add-ons and embedded pieces of software, across multiple infrastructural activities.[249] Therefore, a handful of intermediaries have evolved from single platform services to be infrastructure-like, ubiquitous and essential in the overall advertising and data industry. Consequently, they can significantly profit from disinformation online, including from infrastructural back-ends even if deceptive content is demonetized or ads removed.[250]

3.4. COVID-19 disinformation

The global health crisis has highlighted the importance of information and the need for reliable, accurate journalism.[251] While studies indicate that people trust the online information system less than legacy media,[252] when much of the world turned online, so

[243] Global Disinformation Index (GDI), Popular Brands Appearing Next to COVID-19 Anti-vaccination Disinformation, February 2021; https://disinformationindex.org/wp-content/uploads/2021/02/Feb_11_2021-DisinfoAds-_EU_COVID-19_AntiVaxx.pdf.
[244] GDI, Ad-funded COVID-19 Disinformation: Money, Brands and Tech, July 2020; https://disinformationindex.org/wp-content/uploads/2020/07/GDI_Ad-funded-COVID-19-Disinformation-1.pdf, p 3. These estimates are likely only the tip of the iceberg, see p 19.
[245] GDI, Ad-funded COVID-19 Disinformation, p 4.
[246] Yung Au, Philip N. Howard, Project Ainita, Profiting from the Pandemic: Moderating COVID-19 Lockdown Protest, Scam, and Health Disinformation Websites, COVID-19 Series, Oxford Internet Institute, November 2020; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2020/12/Profiting-from-the-Pandemic-v8-1.pdf.
[247] Au et al, Profiting from the Pandemic, p 5, see also Tech Transparency Project, Google is Paying Creators of Misleading Coronavirus Videos, April 2020; https://www.techtransparencyproject.org/articles/google-paying-creators-of-misleading-coronavirus-videos.
[248] Au et al, Profiting from the Pandemic, p 6.
[249] Au et al, Profiting from the Pandemic, p 6ff.
[250] *Ibid*, p 8.
[251] ICFJ and Tow Center for Digital Journalism at Columbia University, Journalism and the Pandemic.
[252] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 16. For recent data on trust in legacy media, see Reuters, Digital News Report 2021.

did the consumption of information, with trends of data overexposure and declining trust exacerbating.[253]

The early uncertainties and political anxiety during the pandemic significantly increased peoples' susceptibilities to falsehood.[254] The (still) unverified nature of much of the knowledge on the virus and its variants,[255] the evolving data with inconsistencies and constant corrections of the prevalent scientific opinion have paved the way for hoaxes and conspiracies.[256] Although statistics are rare, it is estimated that up to 40% of COVID-19-related social media content comes from unreliable sources, or is even manipulated content.[257]

Early on, with disinformation quickly spreading on the origin and impact of the virus, as well as on health claims, the WHO declared an "infodemic" next to the pandemic.[258] The term refers to an overabundance of information (regardless of its accuracy), which hampers the finding of trustworthy and reliable guidance.[259] In the context of the COVID-19 pandemic, UNESCO identified four main themes of disinformation: (1) emotive narrative constructs, (2) fabricated websites and authoritative identities, (3) fraudulently altered, fabricated, or decontextualized images and videos, and (4) disinformation infiltrators and orchestrated campaigns.[260]

While access to reliable information is crucial at all time, during a health crisis, disinformation diminishes the individual and collective ability to find the best available information to address personal and public health with the potential of devastating consequences, in a worst case scenario, even death.[261] By questioning verifiable information or calling them lies, by misleading people into not believing official sources or independent media guided by professional standards, disinformation weakens those entities, and ultimately attacks public trust.[262] Scientific disinformation often not only dismisses scientific findings but also the process leading to those findings. This reasoning, however, is the basis for informed public policy and good governance.[263]

---

[253] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 18.
[254] Butcher, COVID-19 as a turning point in the fight against disinformation, p 8. Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 370. See also Blackbird, disinformation, p 5. Blackbird.AI and NewsGuard combine AI for an assessment of hoax content at scale.
[255] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 369.
[256] *Ibid*, p 380f. Blackbird.AI, COVID-19 (Coronavirus) Disinformation Report, Volume 4.0, April 2021; https://www.blackbird.ai/reports, p 5.
[257] UNESCO, Journalism, press freedom and COVID-19, p 3 and sources therein.
[258] Philip Howard, Rasmus Kleis Nielson, Nic Newman, J. Scott Brannan, The COVID-19 'infodemic': what does the misinformation landscape look like and how can we respond?, Oxford Internet Institute, April 2020; https://www.oii.ox.ac.uk/blog/the-covid-19-infodemic-what-does-the-misinformation-landscape-look-like-and-how-can-we-respond.
[259] World Health Organization (WHO), 1st WHO Infodemiology Conference, June 2020; https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference.
[260] UNESCO, Disinfodemic 1, p 5.
[261] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 373. In Iran, for example, reportedly over 700 people dies following the spread of false information about drinking high amounts of alcohol to cure the disease. See Al Jazeera, Iran: Over 700 dead after drinking alcohol to cure coronavirus, April 2020; https://www.aljazeera.com/news/2020/4/27/iran-over-700-dead-after-drinking-alcohol-to-cure-coronavirus.
[262] Christopher Dornan, Scientific Disinformation In a Time of Pandemic, Public Policy Forum, June 2020; https://ppforum.ca/publications/science-disinformation-in-a-time-of-pandemic, p 9.
[263] Dornan, Scientific Disinformation In a Time of Pandemic, p 11 and p 15.

Attacking experts and expertise alike,[264] disinformation is often linked to calls to action and resistance, to protest, disobey or even use violence.[265] It can thus additionally threaten public order, for example by provoking attacks against individuals, communities or infrastructures (such as 5G masts).[266]

Disinformation is also linked to decreasing safety of journalists, in particular those reporting from protests against COVID-19 health measures.[267] In short, disinformation hampers the ability of people to make informed decisions and simultaneously attacks democratic values, with the potential to lead to disbelief in any information,[268] to erosion of public trust and the weakening of democratic institutions.[269]

Moreover, disinformation impacts the media landscape also in other ways, as claims of "fake news" enables politicians and other public figures to discredit scrutiny or the verification of statements. The increasing need for fact-checking binds limited resources away from own news investigations, further threatening the sustainability of the already struggling legacy media.[270] Moreover, a struggling media risks being manipulated or accidentally taking over false narratives.[271] In addition, a media heavily relying on vectors such as novelty and conflict might elevate inaccurate hypothesis,[272] or present minority scientific views as equally valid "alternative opinions", resulting in a misleading balance of coverage.[273]

Motivations to spread COVID-19 disinformation can be financial, ideological, or simply reputational to leverage a convenient narrative or distraction. Even if the aim is not to convince about falsehood, disinformation can be utilized to nurture division or erode public trust.[274] Politically motivated disinformation may accept the risk of impeding public health measures,[275] increasing societal tensions,[276] or discrediting scientific information,[277] in order to distract from ineffective COVID-19 responses, for example.

---

[264] ISD, Disinformation Overdose: A study of the Crisis of Trust among Vaccine Sceptics and Anti-Vaxxers, July 2021; https://www.isdglobal.org/isd-publications/disinformation-overdose-a-study-of-the-crisis-of-trust-among-vaccine-sceptics-and-anti-vaxxers, p 28.
[265] ISD, Disinformation Overdose, p 32f.
[266] GDI, Ad-funded COVID-19 Disinformation, p 21. European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 3.
[267] OSCE Representative on Freedom of the Media, Special Report on Handling of the Media During Public Assemblies, October 2020; https://www.osce.org/files/f/documents/2/f/467892.pdf.
[268] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 7.
[269] Reuters, Types, sources, and claims of COVID-19 misinformation.
[270] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 23.
[271] Amel Ghani, Sadaf Khan, Media Matters for Democracy, Disorder in the newsroom: the media's perceptions and response to the infodemic, December 2020;
https://drive.google.com/file/d/1nOgwtFRH5hEtlRBYcayY5aAjEEm8aWBQ/view.
[272] Dornan, Scientific Disinformation In a Time of Pandemic, p 23.
[273] ISD, Disinformation Overdose, p 43. UNESCO, Journalism, 'Fake News' & Disinformation, p 9.
[274] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 8.
[275] Human Rights Council, Forty-first session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "Disease pandemics and the freedom of opinion and expression", David Kaye, A/HRC/44/49, April 2020; https://undocs.org/A/HRC/44/49.
[276] Human Rights Council, Forty-six session, Minority issues, Report of the Special Rapporteur on minority issues, Fernand de Varennes, A/HRC/46/57, March 2021; https://undocs.org/A/HRC/46/57, para 27.
[277] As, for example, also in the context of climate change, see UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 24ff.

In general, disinformation spread by powerful actors can be particularly effective. Early on in the pandemic, falsehood based on state-sponsored blame games proliferated and influencers picked up on them, moving conspiracies from the margins to the mainstream.[278] According to a recent Reuters study, 70% of false COVID-19-related social media content stems from public figures (20% of overall false information).[279] Former U.S. president Donald Trump's suggestion to inject disinfectant,[280] for example, or Madagascar President Andry Rajoelina's proposal to treat COVID-19 with herbal tea may have significantly impaired public health.[281] Other sources that are generally trustworthy can also mislead effectively, in particular if they are using scientific terms or notions.[282] A handful of doctors or nurses, for instance, leveraging the credibility of the medical profession, can create a false sense of uncertainty while in reality there is a clear scientific consensus about health measures or vaccines.[283]

A recent study on Facebook showed that, overall, 12 individuals may be responsible for 65% of COVID-19-related falsehoods on the platform.[284] These "superspreaders" of disinformation are often linked to Facebook pages that are key drivers.[285] Automated content curation may facilitate these dynamics by promoting fringe views or extreme cases, amplifying their reach and consequently intensifying perceived relativism where every opinion can be found online, just as its counterpart.[286] A handful of actors, making use of the digital architecture, could thus trigger significant campaigns against public health measures, such as face masks, social distancing, or lockdowns.[287]

The introduction of COVID-19 vaccinations, in particular, led to an unprecedented number of falsities regarding their safety, efficacy and development.[288] Vaccination disinformation on Facebook increased by 20-30% between April 2020 and April 2021,[289]

---

[278] ISD, Covid-19 Disinformation Briefing No. 1, p 4f. Dornan, Scientific Disinformation In a Time of Pandemic, p 20.
[279] Reuters, Types, sources, and claims of COVID-19 misinformation.
[280] BBC, Coronavirus: Outcry after Trump suggests injecting disinfectant as treatment, April 2020; https://www.bbc.com/news/world-us-canada-52407177.
[281] BBC, Raïssa Ioussouf, Madagascar president's herbal tonic fails to halt Covid-19 spike, August 2020; https://www.bbc.com/news/world-africa-53756752.
[282] Joan Donovan, The Media Manipulation Casebook, Cloaked Science, 2020; https://mediamanipulation.org/definitions/cloaked-science.
[283] NBC, Brandy Zadrozny and Ben Collins, As vaccine mandates spread, protests follow – some spurred by nurses, August 2021; https://www.nbcnews.com/tech/social-media/vaccine-mandates-spread-protests-follow-spurred-nurses-rcna1654.
[284] Center for Countering Digital Hate (CCDH), The Disinformation Dozen, Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers; https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_b7cedc0553604720b7137f8663366ee5.pdf. This is disputed by the company, see Facebook, How We're Taking Action Against Vaccine Misinformation Superspreaders, August 2021; https://about.fb.com/news/2021/08/taking-action-against-vaccine-misinformation-superspreaders.
[285] Avaaz, Facebook's Algorithm: A Major Threat to Public Health, August 2020 https://secure.avaaz.org/campaign/en/facebook_threat_health.
[286] European Parliament, Disinformation and propaganda, p 59.
[287] Au et al, Profiting from the Pandemic, p 1.
[288] The Virality Project, COVID-19 vaccine policies, February 2021; https://www.viralityproject.org/policy-analysis/evaluating-covid-19-vaccine-policies-on-social-media-platforms.
[289] ISD, Disinformation Overdose, p 9.

and so-called "anti-vaxxers" increased their following base by almost 8 million between 2019 and October 2020 (generating over 1 billion USD advertising revenue).[290]

In general, crises such as a pandemic can not only be exploited by criminal actors,[291] but also by extremist movements and hate groups for racist, anti-Semitic or anti-government disinformation.[292] Studies evidences how far-right networks, for example, used the crisis to spread disinformation targeting migrants and promoted the idea that democracy failed to motivate people to advance its end by fueling violence and social conflict.[293] QAnon, for instance, has significantly capitalized on the pandemic and increased their reach online.[294]

Disinformation, moreover, is regularly gendered, and linked to harassment and other forms of violence. Misogynistic narratives have been adapted to the health crisis to undermine women's democratic and digital participation,[295] further entrenching the pandemic's gendered impact due to different access to information, and under-inclusive authoritative facts and voices.[296] Several studies show that COVID-19 disinformation disproportionally affects certain groups in society, with false narratives framing certain identities and strategically targeting marginalized communities.[297]

---

[290] Talha Burki, The online anti-vaccine movement in the age of COVID-19, The Lancet, Digital Health, Volume 2, Issue 10, October 2020; https://doi.org/10.1016/S2589-7500(20)30227-2. This also confirms the effectiveness of vaccination conspiracy groups. Already during the 2019 measles outbreak, anti-vaccination activates were the fasted growing camp on Facebook and successfully targeted undecided individuals. See Dornan, Scientific Disinformation In a Time of Pandemic, p 31f and sources therein.

[291] United Nations Interregional Crime and Justice Research Institute (UNICRI), Stop the virus of disinformation. The risk of malicious use of social media during COVID-19 and the technology options to fight it, November 2020; http://www.unicri.it/sites/default/files/2020-11/SM%20misuse.pdf, p 6f.

[292] ISD, Covid-19 Disinformation Briefing No. 1, p 6f. RAND, Kate Cox, Theodora Ogden, Victoria Jordan, Pauline Paille, COVID-19, Disinformation and Hateful Extremism: Literature review report, prepared for the Commission for Countering Extremism (CCE), March 2021; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993841/RAND_Europe_Final_Report_Hateful_Extremism_During_COVID-19_Final_accessible.pdf.

[293] ISD, Covid-19 Disinformation Briefing No. 1, p 2.

[294] ISD, Covid-19 Disinformation Briefing No. 2, April 2020; https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-2, p 9f.

[295] EU Disinfo Lab, Misogyny and Misinformation: An Analysis of Gendered Disinformation Tactics During the DOCID-19 Pandemic, December 2002; https://www.disinfo.eu/publications/misogyny-and-misinformation%3A-an-analysis-of-gendered-disinformation-tactics-during-the-covid-19-pandemic.

[296] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 102. Nina Jankowicz, How Disinformation Became a New Threat to Women, Female politicians and other high profile women face a growing threat from sexualized disinformation, Coda Story, December 2017; https://www.codastory.com/disinformation/how-disinformation-became-a-new-threat-to-women. Julie Posetti and Kalina Bontcheva, UNESCO, Disinfodemic: Dissecting responses to COVID-19 disinformation, Policy brief 2, May 2020; https://en.unesco.org/sites/default/files/disinfodemic_dissecting_responses_covid19_disinformation.pdf, p 11. More than 40% of attacks against women journalists, for example, are linked to orchestrated disinformation campaigns. See, ICFJ and UNESCO, Global Study: Online Violence Against Women Journalists, November 2020; https://www.icfj.org/our-work/icfj-unesco-global-study-online-violence-against-women-journalist, and in particular, ICFJ, Maria Ressa: Fighting on Onslaught of Online Violence – A big data analysis, March 2021; https://www.icfj.org/our-work/maria-ressa-big-data-analysis.

[297] GDI, Ad-funded COVID-19 Disinformation, p 21. Dhanaraj Thakur, DeVan L. Hankerson, Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender; Center for Democracy & Technology, February 2021; https://cdt.org/wp-content/uploads/2021/02/2021-02-10-CDT-Research-Report-on-Disinfo-Race-and-Gender-FINAL.pdf.

Moreover, as information is unevenly distributed in society, income, education, and online skills become potential barriers to accessing and processing reliable COVID-19 information.[298] Poorly investigated and unverified content is regularly free, which means that individuals who cannot afford quality journalism or lack access to public service media are particularly vulnerable to falsehood.[299] Given the link between disinformation and hygienic rules, people who consume disinformation tend to ignore official health advice and thus engage in unsafe behavior.[300] Consequently, COVID-19 disinformation severely impedes the health of individuals and societies, and entails disproportionate risks for already marginalized groups. Taking this intersectional perspective into account is essential to find sustainable responses to the pandemic and disinformation relating to it.

3.5. How disinformation works on the human level

To comprehend the extent and impact of COVID-19 disinformation, it is necessary to understand underlying human behavioral factors.[301] Disinformation feeds on psychological biases, such as the confirmation bias, describing how people seek out information confirming their existing beliefs and interpret content in such a way.[302] This includes the tendency to ignore inconsistencies that oppose one's own beliefs.[303] People also tend to use motivated reasoning to process information in order to conclude in a way that suits their ideology or some end goal.[304] While algorithms may reduce individuals' exposure to alternative views,[305] people themselves show preference in viewing and sharing of ideologically congenial (dis-)information of demographically similar groups.[306] Thus, falsehood spreads particularly well in trusted networks and by side-stepping legacy media and scrutiny.[307] Moreover, "availability heuristic", describes the phenomenon that people will rather regard something as true they can recall, so continuous exposure to an idea artificially validates it.[308]

[298] Residorf et al., Information Seeking Patterns and COVID-19 in the United States, p 8f.
[299] UNESCO, Journalism, 'Fake News' & Disinformation, p 8.
[300] European Parliament, Disinformation and propaganda: 2021 update, p 101. European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 2.
[301] European Parliament, Study on regulating disinformation with artificial intelligence, p 2.
[302] D.J. Flynn, Brandan Nyhan, Jason Reifler, The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics, Political Psychology, Vol. 38, Issue S1, pp. 127-150, January 2017; https://doi.org/10.1111/pops.12394.
[303] UNESCO, Journalism, 'Fake News' & Disinformation, p 91.
[304] *Ibid*, p 92.
[305] *Ibid*, p 62.
[306] Andy Guess, Kevin Aslett, Joshua Tucker, Richard Bonneau, Jonathan Nagler, Cracking Open the News Feed,: Exploring What U.S. Facebook Users See and Share with Large-Scale Platform Data, Journal of Quantitative Description: Digital Media, Volume 1, April 2021; https://doi.org/10.51685/jqd.2021.006. Ryan J. Gallagher, Larissa Doroshenki, Sarah Shugars, David Lazer, Brooke Foucault Welles, Sustained Online Amplification of COVID-19 Elites in the United States, social media + society, June 2021; https://doi.org/10.1177/20563051211024957.
[307] UNESCO, Journalism, 'Fake News' & Disinformation, p 63.
[308] This effect is further accelerated by the highly personalized online environment, see Blackbird.AI, COVID-19 (Coronavirus) Disinformation Report, p 5. For this reason, uncritical reporting of disinformation can also be counterproductive, see UNESCO, Journalism, 'Fake News' & Disinformation, p 92.

Disinformation also has a "participatory nature".[309] This phenomenon has been illustrated by comparing the viewing of contradictory views online to hearing distant opposition in a sport stadium, while sitting with likeminded individuals, than reading a potentially valid opposing view. This in-group feeling is increased by the perceived connection with one's own online communities. A feeling of belonging may be stronger than facts, which also makes it difficult to pull back or correct falsehoods.[310] The "social noise" model further describes that individuals do not interact with information based solely on their beliefs, but are equally concerned about their relationships with others and their personal image, and may hence act in accordance with perceived expectations of others in their online community.[311] Therefore, social norms may significantly impact the perceived trustworthiness of sources and information.[312]

Another relevant factor is that humans often seek quick and definitive answers,[313] especially at times of heightened uncertainty, attention scarcity and social and economic challenges. Disinformation regularly offers a comprehensible story with simple, plausible answers, which may provide some structure in a time of pandemic-related stress and ambiguity.[314] Research shows that individuals are more susceptible to disinformation when they perceive greater life-related stress, or if they generally distrust the media.[315]

For sustainable responses to disinformation, it is thus essential to consider social and psychological factors. Even if online disinformation could be effectively removed through automated content moderation, this would only bury rather than address the underlying sociopolitical questions.[316]

4. International Human Rights Framework

4.1. Public international law

In the early 20th century, states first addressed false and misleading information through public international law. As a state-centric regulatory system, it can provide legal

[309] European Parliament, Disinformation and propaganda, p 36.
[310] MIT Technology Review, Zeynep Tufekci, How social media took us from Tahrir Square to Donald Trump, August 2018; https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump. UNESCO, Journalism, 'Fake News' & Disinformation, p 63.
[311] Tara Zimmerman, Introducing the Concept of Social Noise, University of North Texas, October 2020; http://hdl.handle.net/2142/108848.
[312] Sara Pluviano, Sergio Della Sala, Caroline Watt, The effects of source expertise and trustworthiness on recollection: the case of vaccine misinformation, Cognitive Processing, Vol. 21, pp. 321-330, April 2020; https://link.springer.com/article/10.1007%2Fs10339-020-00974-8.
[313] Nemr and William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age, p 7 and p 9.
[314] European Parliament, Disinformation and propaganda: 2021 update, p 99f.
[315] *Ibid*, p 101.
[316] Gorwa et al., Algorithmic content moderation, p 12. Psychological means could also be used to minimize the creation and spread of disinformation in the first place, for example by introducing reputation systems. For assessments of alternative content governance approaches with a focus on community-led governance models, see Ben Wagner, Johanne Kübler, Eliška Pírková, Rita Gsenger, Carolina Ferro, Reimagining Content Moderation and Safeguarding Fundamental Rights: A Study on Community-Led Platforms, European Parliament Greens/EFA study, May 2021; https://enabling-digital.eu/wp-content/uploads/2021/07/Alternative-content_web.pdf.

tools to react to disinformation spread by state actors.[317] In 1936, for example, the League of Nations drafted the International Convention Concerning the Use of Broadcasting in the Cause of Peace, which prohibits states from transmitting incorrect statements. In 1953, the United Nations Convention on the International Right of Correction was adopted, which entitles states with a right of reply to news dispatches they consider false. Generally, if false government statements intervene in the affairs of another state, the latter can invoke a violation of the principle of non-intervention.[318]

4.2. Human rights law and business and human rights

While various legislation contain some rules concerning falsehood, for example in the context of trade, advertising or criminal fraud, international human rights law provides the legal framework to assess the legitimacy of restrictions to one's freedom of expressions, including deceptive speech. Human rights law is thus also a benchmark for assessing legislation that addresses false information spread through mass media. To promote ensuring accuracy of expression with a wide reach or audience, various national media laws stipulate a right of correction or reply,[319] and provide for self-regulatory professional codes, which typically require journalists to verify information, but also to act as safeguards for reporting in good faith.[320]

While disinformation as a form of expression benefits from some protection, as outlined below, it simultaneously affects the right to freedom of thought, opinion and expression of others. It may also affect the right to non-discrimination, private life, economic and social rights, and, especially in the context of the COVID-19 pandemic, the right to life and health. Disinformation can generally weaken human autonomy, which is at the core of all human rights. And while some human rights may be restricted under certain preconditions, others, such as the right to freedom of opinion and the right to life, are absolute rights not permitting any interference.[321]

As international law, human rights are state-centered, traditionally understood as a protective shield against state interference.[322] States are the primary duty bearer to protect, promote and respect human rights. They thus have a duty to refrain from (unjustified) interferences, and at the same time have a positive obligation to promote and fulfil human rights, and to ensure that others, including private actors, do not infringe them.[323] States' obligations extend to the protection of human rights at the individual as

---

[317] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 9f.

[318] For more information on the historical background, see OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 4. See also Richter, Fake News and Freedom of the Media and RFoM policy brief 1, p 9f.

[319] This right is typically available for statements of facts, while not for value judgments as they cannot be proven to be true or false, see Richter, Fake News and Freedom of the Media, p 24. For further information and assessment of national, regional and international rights of correction or reply, see Richter, Fake News and Freedom of the Media, p 14ff.

[320] UNESCO, Journalism, 'Fake News' & Disinformation, p 96ff.

[321] Walter Berka, Christina Binder, Benjamin Kneihs, Die Grundrechte, Grund- und Menschenrechte in Österreich, 2. Auflage, Verlag Österreich, November 2019, p 243.

[322] Council of Europe, Algorithms and Human Rights, p 33.

[323] Berka et al, Grund- und Menschenrechte in Österreich, p 147f.

well as the collective, societal level.[324] Moreover, the international community recognized that the same rights people have offline are equally protected online.[325]

Relevant to online disinformation, in addition to human rights, the Sustainable Development Goals also aim to ensure "public access to information and protect fundamental freedoms" (SDG Target 16.10), calling on states to "ensure public access to information". Moreover, SDG 3 aims to "ensure healthy lives and promote well-being for all at all ages". [326]

Over the course of the last century, the increasing power and control of private, profit-oriented actors has become increasingly evidently relevant for the fulfilment of human rights. Recognizing this has led to global initiatives such as the United Nations (UN) Global Compact as a platform for promoting corporate social responsibility, and the UN Guiding Principles on Business and Human Rights (UNGP).[327] The UNGP stipulate a responsibility of corporations to respect human rights. This responsibility exists "independently of states' abilities and/or willingness to fulfil their own human rights obligations".[328] While acknowledging that corporations should comply with national law, the UNGP require private actors to honor the principles of internationally recognized human rights, to interpret restrictive laws narrowly and even to challenge them where appropriate.[329] Over the past years, intermediaries have repeatedly been urged to uphold this responsibility to respect human rights, including by the international free speech mandate holders (the UN Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information).[330]

4.3. Right to freedom of opinion and expression

The international human rights framework acknowledges that freedom of opinion and expression are gatekeeper rights as they are "indispensable conditions for the full

[324] European Parliament, Disinformation and propaganda: 2021 update, p 114.
[325] Human Rights Council, Twentieth session, Resolution adopted by the Human Rights Council on 16 July 2012, The promotion, protection and enjoyment of human rights on the Internet, July 2012, A/HRC/RES/20/8; https://undocs.org/A/HRC/RES/20/8 and subsequent decisions; also confirmed by the UNGA, and UN Special Rapporteur on freedom of expression, A/HRC/32/38, para 6.
[325] United Nations, Sustainable Development Goals – Transforming our world: the 2030 Agenda for Sustainable Development, https://sdgs.un.org/2030agenda.
[327] For more information on the UNGP, see chapter 5.3.2.
[328] Principle 11.
[329] See Principle 23 and Global Network Initiative (GNI), Principles on Freedom of Expression and Privacy; https://globalnetworkinitiative.org/gni-principles; and GNI, Implementation Guidelines; https://globalnetworkinitiative.org/implementation-guidelines.
[330] See, *inter alia*, UN Special Rapporteur on freedom of expression, A/HRC/38/35; Joint Declaration on Freedom of Expression and "Fake News"; United Nations General Assembly, Seventy-fourth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "hate speech", David Kaye, A/74/486, October 2019; https://undocs.org/en/A/74/486.

development of the person". Freedom of opinion and expression are "essential for any society" as they "constitute the foundation stone for every free and democratic society".[331]

This paper specifically focuses on freedom of expression, i.e., the right to seek, receive and impart information, as enshrined (with some deviations) in several international, regional, and national human rights frameworks. Most importantly, free speech rights are protected by Article 19 of the Universal Declaration of Human Rights, which, while not being legally binding per se, constitutes customary international law,[332] and Article 19 of the International Covenant on Civil and Political Rights (ICCPR). In the European context, Article 10 of the European Convention on Human Rights (Council of Europe), and Article 11 of the Charter of Fundamental Rights of the European Union guarantee freedom of expression. Similar protections can also be found in the American Convention on Human Rights, the African Charter on Human and Peoples' Rights, the Association of Southeast Asian Nations' Human Rights Declaration and other regional instruments.

Article 19 of the ICCPR includes the absolute right to freedom of opinion and the right to freedom of expression, also encompassing media freedom and access to information. Article 19 hence includes an internal (holding opinions) and external dimension (expression), and the right to express opinions (actively) and to receive information (passively).[333] Freedom of expression is considered as universal "everyone right",[334] protecting any speech regardless of its content, form, or medium used,[335] and regardless of frontiers.[336]

The right to freedom of expression includes all kinds of information and ideas, including those that "shock, offend or disturb".[337] While free speech rights do not include a right to truthfulness, they provide for the right to be informed based on objective factual information, to be able to hold an undistorted opinion,[338] and to obtain information held by public bodies.[339]

States' positive obligation to protect the right to freedom of expression also includes the positive obligation to promote, protect and support diverse and independent

---

[331] Human Rights Committee of the International Covenant on Civil and Political Rights, 102nd session, General comment No. 34, CCPR/C/GC/34, https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf, para 2.
[332] United Nations Office of the High Commissioner for Human Rights, The European Union and International Human Rights Law, https://europe.ohchr.org/Documents/Publications/EU_and_International_Law.pdf. Berka et al, Grund- und Menschenrechte in Österreich, p 24.
[333] Berka et al, Grund- und Menschenrechte in Österreich, p 659.
[334] European Court of Human Rights, Autronic AG v. Switzerland, application no. 12726/87, judgment, 22 May 1990; http://hudoc.echr.coe.int/eng?i=001-57630, para 47.
[335] The right to seek, receive and impart "information and ideas of all kinds". See also European Court of Human Rights, Guide on Article 10 of the European Convention on Human Rights, Freedom of Expression, December 2020; https://www.echr.coe.int/documents/guide_art_10_eng.pdf. Dominiko Bychawska-Siniarska, Council of Europe, Protecting the Right to Freedom of Expression Under the European Convention on Human Rights, A handbook for legal practitioners, July 2017; https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814.
[336] It is thus also relevant for speech restrictions outside the own jurisdiction or with effects abroad. See Joint Declaration on Freedom of Expression and "Fake News", para 1(c).
[337] Human Rights Committee, General comment No. 34, para 47. European Court of Human Rights, Handyside v. the United Kingdom, application no. 5493/72, judgment, 7 December 1976; http://hudoc.echr.coe.int/eng?i=001-57499, para 49.
[338] Sciortino, Fake News and Infodemia at the Time of Covid-19, p 40.
[339] Human Rights Committee, General comment No. 34, para 19.

media.[340] As the free flow of information is essential to ensure freedom of expression, the promotion of pluralism and diversity of sources of information and the proactive disclosure of information of public interest held by authorities are considered indispensable.[341]

4.4. Speech restrictions

Freedom of expression is no absolute right. Certain restrictions of the right to seek, receive and impart information can be justified under international human rights law. Article 19(3) outlines the cumulative requirements for such justifications, namely that the restriction is to be provided by law and necessary for respecting the rights and reputations of others, or for protecting national security, public order (ordre public) or public health or morals.

The first criterion, legality, requires restrictions to have a legal basis and that the scope and meaning as well as effect of such law is sufficiently clear, precise and publicly available (to ensure foreseeability). Clear legal criteria must enable an objective assessment, which can be interpreted by independent judicial authorities.[342] The second criterion, legitimacy, refers to the six listed aims that can legitimize restrictions, with falsity not being among them (despite having been considered during the drafting process of the ICCPR[343]). Necessity, as the third criterion, requires proportionality and a causal relationship between the speech to be restricted and the harm to be prevented indicating a substantial link to the need for protection. Restrictive measures hence need to be appropriate and proportionate to achieve a legitimate aim, based on the severity and immediacy of the specific threat, using the least intrusive means.[344] As the human rights framework particularly protects information of public interest and public figures,[345] restrictions referring to such speech need to be particularly narrow, time-limited and tailored to be proportionate.[346]

On the other hand, international human rights law also acknowledges that certain expressions can limit the freedom of expression of others, especially marginalized voices. Consequently, certain expressions require a prohibition. Article 20(2) of the ICCPR, for example, stipulates that "any advocacy of national, racial or religious hatred that constitutes incitement of discrimination, hostility or violence shall be prohibited by law", with the Rabat Plan of Action providing an authoritative roadmap for its implementation.[347] Other international frameworks also require certain prohibitions, for

[340] Joint Declaration on Freedom of Expression and "Fake News", pp 9.
[341] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 38.
[342] Human Rights Committee, General comment No. 34, para 34.
[343] The ICCPR Drafting Committee discussed the possibility to include that speech restrictions may be justified if a "systemic diffusion of deliberately false or distorted reports […] undermine[s] friendly relations between peoples and States". See Richter, Fake News and Freedom of the Media, p 11f and sources therein.
[344] Human Rights Committee, General comment No. 34, para 34.
[345] *Ibid*, para 38.
[346] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 42.
[347] See Human Rights Council, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4; Appendix, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, January 2013; https://undocs.org/A/HRC/22/17/Add.4. The Rabat Plan of

example Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination, Article 3 of the Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography, or Security Council Resolution 1624 (2005) regarding incitement to commit a terrorist act.

4.5. Disinformation and Freedom of Expression

In principle, the expression of false information is protected by international human rights law, albeit with some exceptions. If public discourse is dominated by falsities, however, the primary purpose of free speech is undermined,[348] which risks to reinforce existing divisions in society, fuel hate speech,[349] and result in violence.[350] Freedom of expression, thus, is both the objective and means for combating disinformation.[351]

Consequently, states' positive obligation to protect, respect and promote the right to freedom of expression necessitate certain measures to address disinformation. Responses themselves, however, need to be rooted in international human rights law, as affirmed by the United Nations Human Rights Council. Any restriction on freedom of expression must be based on the three-part test of legality, legitimacy, and necessity and proportionality,[352] with responses tailored to whether discrimination or violence is incited, and whether speech is illegal or legal (yet harmful).

In general, prohibitions of specific speech based on vague or ambiguous notions are not compatible with international law,[353] in particular if they also target the media. Speech restrictions, in principle, cannot be justified if they interfere with the right and role of journalists to impart information of public interest.[354] While falsity alone is no legitimate ground for restrictions, legal criteria around intent to deceive are equally elusive and hence difficult. Consequently, many regulations focus on the impact of disinformation,[355] or compliance with due diligence duties such as in the journalism context.[356]

The European Court of Human Rights (ECtHR) underlines that any restriction on speech must correspond to a "pressing social need". While there is limited case law regarding restrictions based on falsity,[357] the ECtHR recognizes that speech prohibitions, even with a strong suspicion that the information is not truthful, would deprive

---

Action identifies six factors to determine the severity and necessary to criminalize incitement: context, status of the speaker, intent, content and form of speech, reach of the speech, and likelihood of risk.

[348] European Parliament, Disinformation and propaganda, p 79.

[349] Access Now, Fighting Misinformation and Defending Free Expression During COVID-19: Recommendations for States, April 2020; https://www.accessnow.org/cms/assets/uploads/2020/04/Fighting-misinformation-and-defending-free-expression-during-COVID-19-recommendations-for-states-1.pdf, p 20.

[350] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 15.

[351] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 83

[352] Human Rights Council, Forty-fourth session, Resolution adopted by the Human Rights Council on 16 July 2020, Freedom of opinion and expression, A/HRC/44/12, July 2020; https://undocs.org/en/A/HRC/RES/44/12, pp 12.

[353] Joint Declaration on Freedom of Expression and "Fake News", para 2(a).

[354] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 6.

[355] European Parliament, Disinformation and propaganda: 2021 update, p 36.

[356] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 5.

[357] For an assessment of the existing case law, see RFoM policy brief 1, p 20ff.

expression.[358] The ECtHR also states that the dangers inherent to prior restraints call for "the most careful scrutiny", and any content blocking requires "effective judicial review to prevent any abuse of power".[359] This hesitance to recognize justified restrictions of false information is also reflected in decisions related to election-related disinformation laws, where the ECtHR found violations of Article 10 of the European Convention on Human Rights due to a lack of safeguards when identifying "untrue information" or correlating harms to other human rights.[360]

To assess the proportionality of speech restrictions, the distinction between statements of fact and value judgments is relevant. In the context of disinformation, this differentiation is sometimes difficult. The ECtHR stipulates that "the proportionality of an interference may depend on whether there exists a sufficient factual basis for the impugned statement. […] The existence of facts can be demonstrated, whereas the truth of value judgments is not susceptible of proof. The requirement to prove the truth of a value judgment is impossible to fulfil and infringes freedom of opinion itself".[361]

Furthermore, proportionality requires a clear cause-and-effect relationship between the specific disinformation and its social effect, which is regularly difficult to establish.[362] In this context, it is also necessary to assess whether restrictive measures fulfil their protective function. Restricting online disinformation, however, may not result in the belief of facts or limit harm, which would be necessary to ensure proportionality of restrictions. On the contrary, the removal of disinformation can in some instances even strengthen the original viewpoint or push certain groups out of mainstream information channels to more obscure and less moderated niche platforms.[363] De-platformizations, account blockings, or other systematic removals may drive networks underground, depriving them of critical sources of information.[364] Such trends have been identified in the context of radicalization, where studies show restrictive measures may have cultivated exactly such attitudes they aimed to prevent.[365]

---

[358] European Court of Human Rights, Salov v. Ukraine, application no. 65518/01, judgment, 6 September 2005; http://hudoc.echr.coe.int/eng?i=001-70096, para 113.

[359] European Court of Human Rights, Ahmet Yildirim v Turkey, application no. 3111/10, judgement, 18 December 2012; http://hudoc.echr.coe.int/eng?i=001-115705, para 64.

[360] Joris van Hoboken, Ronan Ó Fathaigh, Regulating Disinformation in Europe: Implications for Speech and Privacy, UC Irvine Journal of International, Transnational, and Comparative Law, Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech, Article 3, May 2021; https://scholarship.law.uci.edu/ucijil/vol6/iss1/3, p 24f and judgments cited therein, especially European Court of Human Rights, Kwiecien v. Poland, application no. 51744/99, judgment, 9 January 2007; http://hudoc.echr.coe.int/eng?i=001-115705.

[361] Settled case-law by the ECtHR since European Court of Human Rights, Lingens v. Austria, application no. 9815/82, judgment, 8 July 1986; http://hudoc.echr.coe.int/eng?i=001-57523, para 46.

[362] European Parliament, Disinformation and propaganda: 2021 update, p 37.

[363] For example to Parler, a platform differentiating its service from other online platforms on the basis of moderating less. This, in turn, increasingly drives the question of moderation to the infrastructural level. App store providers for iOS and Android, for example, were confronted to consider whether to deny Parler access to thier stores. For more on infrastructural moderation, see Jonathan Zittrain, The Inexorable Push for Infrastructure Moderation from the *it's-coming-whether-we-like-it-or-not* dept, Tech Policy Greenhouse by Techdirt, September 2021; https://www.techdirt.com/articles/20210924/12012347622/inexorable-push-infrastructure-moderation.shtml.

[364] Bloch, Automation in Moderation, p 79.

[365] Keller, Internet Platforms, Observations on Speech, Danger, and Money, National Security, Technology, and Law, p 22ff.

Consequently, restrictions of speech entailing falsehood that traceably causes significant harm to public health can be justified. For example, human rights may justify certain actions against anti-vaccination content.[366] It is more difficult, however, to justify a restriction for misleading statements based on value judgments, on disputed factual bases, or if information is presented in a manner that makes it likely to draw false conclusions while the stated facts themselves are true. The presentation and framing of facts are value judgments, so their truthfulness is not susceptible to proof, even if the framing is intentionally misleading.[367]

As it proves difficult to justify restraints on expressing false information, some restrictions focus on the dissemination of false statements. They could be proportional if needed to protect the right to health or the right to private life, for example, or also to tackle aggressive data-driven targeting methods of online dissemination of falsehood.[368]

Any speech restriction focusing on the falsity of content alone need to account for difficult context assessments.[369] According to the ECtHR, the only justifiable restriction of falsehood without regard to context is in the context of Holocaust denial.[370]

4.6. Freedom of opinion and other human rights affected by disinformation

While this paper focuses on the right to freedom of expression, disinformation can affect several other human rights. Article 12 of the International Covenant on Economic, Social and Cultural Rights (ICESCR), for example, provides for the right to health as also enshrined in the Universal Declaration of Human Rights. The UN Committee on Economic, Social and Cultural Rights (ICESCR) emphasizes that "information accessibility" is a key component of the right to health.[371] The ICESCR also includes the right to participation in cultural life, membership of a community and education.

Disinformation can also impact the right to protection of honor and reputation as enshrined in Article 17 of the ICCPR or to the right to be free from discrimination in Article 2 and Article 25 of the ICCPR. Moreover, in the context of restrictions, the right to due process (Article 6 ECHR) and the right to an effective remedy (Article 13 ECHR) are essential. They require states to ensure judicial independence and an effective complaint mechanism that promptly remedies the grievances of individuals, also if private actors caused the interferences.[372] Redress opportunities should be known and accessible to

---

[366] David Kaye, The Clash Over Regulating Online Speech, SLATE, June 2019; https://slate.com/technology/2019/06/social-media-companies-online-speech-america-europe-world.html.

[367] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 5

[368] European Parliament, Disinformation and propaganda: 2021 update, p 38

[369] *Ibid*, p 37.

[370] Speech restrictions for false information about the Armenian Genocide, for example, would require to include context as it is not an as widely accepted historical fact. See European Parliament, Disinformation and propaganda: 2021 update, p 37.

[371] Committee on Economic, Social and Cultural Rights, Twenty-second Session, General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12), E/C.12/2005/4, https://www2.ohchr.org/english/bodies/cescr/docs/gc14.

[372] Council of Europe, Algorithms and Human Rights, p 23f.

everyone, and be prompt, based on individual notice, and provide for a thorough and impartial investigation.[373]

Disinformation may also interfere with the very right to freedom of opinion. As an absolute right, freedom of opinion does not permit any interference or restriction. It encompasses protection against the disclosure of one's opinion, the manipulation in the forming and holding of opinions, and the penalization of one's opinion.[374] It also includes the right not to express one's opinion and the right to change it (including through the influence of others).[375]

While protecting against manipulation, freedom of opinion acknowledges that a constant exposure to a wide range of influence is part of human autonomy and a basis for the formation of one's opinion.[376] Any influence in the thinking process, however, should be based on knowledge and consent of the rightholder states the UN Special Rapporteur on freedom of expression. [377] Brainwashing, for example, constitutes a clear violation of freedom of opinion, as it interferes with the right to form and develop an opinion by way of reasoning.[378] Also other coercive or opaque manipulative techniques may interfere with the right to form and hold opinions.[379]

Consequently, today's information landscape with massive and subliminal influences beyond individuals' knowledge or consent in itself affects the right to form and hold opinions, regardless of whether or not falsehood is involved. The contemporary digital ecosystem with large intermediaries' business practices arguably undermines mental autonomy.[380] The constant and systemic collection and analysis of most personal information to monetize it, optimize emotional engagement, and capture individuals' attention raises questions of coercion,[381] potentially rising to an unacceptable level of persuasion that interferes with the very essence of freedom of opinion. Moreover, recording individuals' most private thoughts enables exploitation and discrimination[382] and encourages addictive engagement, which further undermines user agency and choice.[383]

---

[373] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 14f.
[374] Evelyn Aswad, Losing the Freedom to Be Human, Columbia Human Rights Law Review, Vol. 53, February 2020; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635701, p 307.
[375] Human Rights Committee, General comment No. 34, para 9.
[376] Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, p 33.
[377] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 34. The UN Special Rapporteur underlines that the framing of individuals as "user" by companies distracts from the broader impact both individually and on the societal level, which requires a broader approach than "user concerns".
[378] Manfred Nowak, UN Covenant on Civil and Political Rights: CCPR Commentary, 2nd Edition, 2005.
[379] Susie Alegre, Rethinking Freedom of Thought for the 21st Century, Doughty Street Chambers, April 2020, European Human Rights Law Review, 2017, Issue 3; https://susiealegre.com/wp-content/uploads/2020/04/Alegre%20from%202017_EHRLR_Issue_3_Print_final_0806%5B6745%5D.pdf. Aswad, Losing the Freedom to Be Human. UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 10.
[380] Aswad, Losing the Freedom to Be Human; Alegre, Freedom of Thought, UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 66.
[381] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 11.
[382] Aswad, Losing the Freedom to Be Human.
[383] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 66f.

In this context, Article 12 of the UDHR and Article 17 of the ICCPR are also relevant, stipulating that no one shall be subjected to arbitrary or unlawful interference with their privacy. Article 8 of the ECHR and Article 7 of the EU Charter of Fundamental Rights also protect privacy. According to the UN High Commissioner for Human Rights (OHCHR), the right to privacy includes that the processing of personal data must be "fair, lawful and transparent".[384] Data protection laws, in particular, refer to the processing of personal data that identifies individuals or makes them identifiable. Such laws typically require transparency and a legal basis for data processing, such as consent, and provide for the right to access, correction, deletion and independent regulatory oversight. Moreover, they regularly include specific rules for automated decision-making. The General Data Protection Regulation (GDPR) of the European Union, for example, ties the legitimacy of fully automated decisions (including profiling) to the explicit, informed and genuinely free consent of the affected individual if the automated decision "produces legal effects" or "similarly significantly affects" them (Article 22).[385] This provision may also apply to micro-targeting aimed at influencing behavior.[386] Arguably, automated content governance without meaningful human intervention can be understood as causing effects similar to legal consequences given their far-reaching impact on the realization of freedom of expression.

5. Responses to COVID-19 online disinformation

5.1. Introduction

5.1.1 Outline

Since the outbreak of the COVID-19 pandemic, governments have adopted a variety of regulatory approaches to address falsities and intermediaries have taken unmatched steps to moderate content on their services and to promote official sources of information. While, generally, responses to COVID-19 online disinformation ground in the genuine imperative to address disinformation for the sake of public health, controlling information streams whilst weakening scrutiny is also a temptation for states and intermediaries alike.[387] Censorship, however, just as disinformation, undermines freedom of expression and public health.

UNESCO identified four categories of disinformation responses in the context of the COVID-19 pandemic: responses aimed at (1) the identification of deceptive content such as fact checking or investigative responses by the media, academia, civil society, etc.; (2) the creators and disseminators such as counter-narratives and policy responses;

---

[384] Human Rights Council, Thirty-ninth session, Report of the United Nations High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/39/29, August 2018; https://undocs.org/A/HRC/39/29, para 29.

[385] It may also be legitimate if authorized by law, necessary for the preparation and execution of a contract, provided there are sufficient safeguards in place, which includes informing the affected individual and a possibility to contest the decision, and protections from discrimination based on protected criteria as race, gender, political opinion etc. For more information, see Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, February 2018; https://ec.europa.eu/newsroom/article29/items/612053/en.

[386] European Parliament, Disinformation and propaganda, p 776.

[387] Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, p 3.

(3) the production and distribution mechanisms such as curatorial responses or technical ones as de-monetization; and (4) the target audience such as educational responses, empowerment efforts, and efforts to promote media pluralism.[388]

In general, both states and intermediaries, whose responses are often closely intertwined, searched for technical solutions. Yet, societal challenges such as disinformation cannot be solved merely by outsourcing processes and decision-making to automation.[389] Nevertheless, the pandemic has resulted in a "massive experiment" in automated content moderation, also driven by social distancing requirements and human moderators being sent home, which led to humans being taken out of the loop, leading to dramatically higher numbers of takedowns with, as initial analyses show, questionable accuracy rates[390] and erroneous news content censoring.[391]

5.1.2 Fact-checking

Ever since the beginning of the pandemic, several actors, including civil society, have provided fact-checking, shared data, and developed tools to investigate falsities and deception.[392] Most intermediaries set up or increased cooperation with independent fact-checking organizations and trusted flaggers. Despite a significantly increased collaboration, intermediaries have been frequently accused of not sufficiently empowering fact-checkers with data access and due prominence to their scrutinized content.[393] Meanwhile, individuals and users are typically not enabled to report on or correct falsehood.[394]

While fact-checking has been an important tool, it struggles to eliminate the adverse impact of disinformation[395] as it rarely reaches the same audience as online falsities,[396] and is typically a slow procedure. As it is time-consuming, costly, and difficult to scale, fact-checking binds extensive resources.[397] While fact-checkers' efficacy and

---

[388] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Julie Posetti and Kalina Bontcheva, UNESCO, Disinfodemic: Deciphering COVID-19 disinformation, Policy brief 1, April 2020; https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf, p 7.

[389] Gorwa et al., Algorithmic content moderation, p 12.

[390] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 10 and sources therein.

[391] The Verge, Facebook was marking legitimate news articles about the coronavirus as spam due to a software bug, March 2020; https://www.theverge.com/2020/3/17/21184445/facebook-marking-coronavirus-posts-spam-misinformation-covid-19.

[392] See, for example, #CoronaVirusFacts Alliance, Poynter, https://www.poynter.org/coronavirusfactsalliance. For collaboration and investigation tools, see Bellingcat, Investigating Coronavirus Fakes And Disinfo? Here Are Some Tools For You, March 2020; https://www.bellingcat.com/resources/2020/03/27/investigating-coronavirus-fakes-and-disinfo-here-are-some-tools-for-you. For data sharing see Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 379.

[393] European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 8.

[394] Council of Europe, Parliamentary Assembly, Resolution 2143.

[395] Man-piu Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, Dolores Albarracin, Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation, Psychological Science, Vol. 28, Issue 11, pp. 1531-1546, May 2017; https://doi.org/10.1177/0956797617714579.

[396] In particular due to complex like and sharing infrastructural settings, see Gray, Fake news, p 328 and 330 figure 8.

[397] European Parliament, Disinformation and propaganda, p 114.

credibility is essential for the process,[398] they may themselves face selection biases or struggle with complex scientific contexts.[399] The COVID-19 pandemic has thus generated an enormous challenge even for well-established fact-checkers.[400]

Responses following the debunking of content include depriorization, warning labels, and removals. Adding warnings to fact-checked content, for example, that differentiate between "disputed" and "rated false", is among the least intrusive measures. However, studies have indicated that disclaimers of potential manipulation only have limited effects.[401] Furthermore, studies show that such efforts should be narrow in order to avoid spillovers that could result in an overall decreased belief in information.[402]

Lastly, it is important to recognize that while exposing individuals to countervailing information can be successful, the person may still be reluctant to adjust their beliefs.[403] Although research remains tenuous, there is some evidence that corrective measures may even be counterproductive by drawing more attention to the questionable content.[404] Removals or de-platforming could be understood as a badge of honor, or act as "proof" of a conspiracy or authorities' claimed attempts to silence.[405]

5.1.3 International responses

Recognizing the severity and cross-border nature of COVID-19 disinformation, several initiatives to tackle it were undertaken at the international level. The World Health Organization (WHO) swiftly initiated the Information Network for Epidemics (EPI-WIN) aimed at ensuring veracity of information, bringing together various actors to amplify accurate information, and to respond to falsehoods through "myth-busting". In addition, the WHO introduced the Risk Communication and Community Engagement (RCCE) and partnered up with several internet intermediaries to promote authoritative content. The United Nations additionally launched an SOS Alert in all official UN languages, closely cooperating with internet intermediaries to target specific populations and demographics with information (especially with Facebook and Instagram) and to

[398] European Parliament, Disinformation and propaganda: 2021 update, p 62f.
[399] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 383 and p 385.
[400] Eduardo Suárez, Reuters, How fact-checkers are fighting coronavirus misinformation worldwide March 2020; https://reutersinstitute.politics.ox.ac.uk/risj-review/how-fact-checkers-are-fighting-coronavirus-misinformation-worldwide.
[401] European Parliament, Disinformation and propaganda: 2021 update, p 102 - studies show that content involving critical comments are rather not shared than those including warning labels.
[402] Katherine Clayton, Spencer Blair, Jonathan A. Busam et al, Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media, Political Behavior, Vol. 42, Issue 4, pp 1073-1095, December 2020; https://doi.org/10.1007/s11109-019-09533-0, p 1091.
[403] Nuñez, Disinformation Legislation and Freedom of Expression, p 788.
[404] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 22. Thomas Wood, Ethan Porter, The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence, Political Behavior, Volume 41, Issue 1, pp 135-163, March 2019; https://doi.org/10.1007/s11109-018-9443-y.
[405] Ofcom, Use of AI in Online Content Moderation, p 44. Dornan, Scientific Disinformation In a Time of Pandemic, p 21.

promote the visibility of verified information (especially with Google).[406] UNESCO established a Resource Center of Responses to COVID-19.[407]

On a regional level, several supranational and regional organizations undertook efforts to combat COVID-19 disinformation through existing mechanisms and newly introduced approaches. The European Union deployed its Rapid Alerts System, established by the 2018 Action Plan Against Disinformation, to enable the exchange of information and expose disinformation in real time, building on experience of its East StratCom Task Force and EUvsDisinfo project.[408] Overall, EU responses vary depending on whether false information is spread intentionally and whether influence operations are orchestrated by third country actors. Responses range from targeted rebuttals to literacy efforts and coordinated government actions.[409] The EU also provides an overview of false narratives[410] and undertakes various initiatives to support democratic infrastructures and strengthen media freedom, for example with the Media4Democracy programme, the media literacy week, and the European Democracy Action Plan. The EU aims at promoting literacy to increase individuals' ability to discern the quality of information and to navigate today's complex media ecosystem.[411] As other international actors, the EU underlines the crucial role of media freedom for reliable and fact-checked information, as well as for scrutinizing and ensuring accountability for state responses to the pandemic.[412] The EU Observatory for Social Media Analysis (SOMA) and the European Digital Media Observatory (EDMO) focus on strengthening journalism, supporting fact-checking and building resilience and independent governance.[413]

An important tool in the EU's responses to COVID-19 disinformation is its 2018 Code of Practice, a voluntary, self-regulatory mechanism that has been agreed on by online platforms, advertisers and the advertising industry ("induced self-regulation"[414]). The Code requires scrutiny of ad placements to incentivize against profiting from falsehood and informing users why they are targeted by specific advertisements, who is the sponsor, and what is the amount paid. The Code includes rules against fake accounts or the use of automated bots, and on the empowerment of users by requesting intermediaries to provide tools to find diverse perspectives about topics of public

[406] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 374f and WHO's responses are summarized in the WHO timeline, see https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline.
[407] UNESCO, Resource Center of Responses to COVID-19, https://en.unesco.org/covid19/communicationinformationresponse/mediasupport.
[408] EUvsDisinfo also includes an overview of corona-related reports, see EUvsDisinfo, https://euvsdisinfo.eu/category/blog/coronavirus.
[409] European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 4.
[410] Via EUvsDisinfo: For an assessment regarding the vaccine rollout and associated disinformation, see, for example, EUvsDisinfo, Dizzy by Vaccine Disinfo - Capturing Vaccines Rollout in The EU's Neighbourhood and Russia, March 2021 https://euvsdisinfo.eu/dizzy-by-vaccine-disinfo-capturing-vaccines-rollout-in-the-eus-neighbourhood-and-russia. See also Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 377f.
[411] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 21.
[412] European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 10.
[413] For more information on EU disinformation initiatives, with a focus on external actions, see European Parliament, The impact of disinformation on democratic processes and human rights in the world, pp 29-38.
[414] Since the Code of Practice was initiated by the European Commission, it is not entirely self-regulatory. See European Parliament, Disinformation and propaganda, p 105.

interest, and to improve the findability of trustworthy content. It also includes rules on digital literacy, partnerships with civil society and the research community by requested access to data, and promoting independent fact-checkers. Moreover, the Code provides transparency for users on general information on algorithms. In order to ensure its implementation, the Code requires a periodic monitoring with publicly available self-assessment reports.[415]

While the Code is not legally binding, it aims to coopt private actors to enforce the public policy objectives of addressing disinformation, in particular by allowing – or even incentivizing – certain restrictions on lawful speech.[416] Although this aims at improving the online information landscape, it inevitably also leads to the sidestepping of certain safeguards that the legal system normally grants for restrictions of speech introduced by state actors.[417] While the Code has generally been recognized as important, it has also been criticized for falling short of holding intermediaries accountable[418] due to the lack of mandating appeals,[419] and for not sufficiently engaging the ad sector.[420] Various actors, moreover, called for more transparency regarding its implementation and oversight.[421] Starting in 2021, the Code of Practice on Disinformation is being updated.[422]

Since a few months after the start of the pandemic, the European Commission requires periodic reporting on initiatives to promote authoritative content, to improve users' awareness, and to address manipulative behavior, and provide information on advertising linked to COVID-19 disinformation.[423]

Other EU legal frameworks focus on processes rather than content itself. The Digital Services and Markets Act package will introduce systemic rules for online platforms and aims to provide for transparency and due process to better protect human rights and fair competition, and ensure greater public accountability.[424] Overall, the Digital Services Act (DSA) constitutes a step towards treating very large internet intermediaries' services as

---

[415] For an assessment of the EU Code of Practice on disinformation, see Aleksandra Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, Disinformation and Digital Media as a Challenge for Democracy, European Integration and Democracy Series, pp. 291 – 308, Intersentia, Vol. 6, June 2020; https://doi.org/10.1017/9781839700422.017, pp 3ff.

[416] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 32.

[417] Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, p 11.

[418] European Court of Auditors, Special Report on Disinformation affecting the EU: tackled but not tamed, June 2021; https://www.eca.europa.eu/Lists/ECADocuments/SR21_09/SR_Disinformation_EN.pdf, p 5.

[419] Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, p 9f.

[420] GDI, Research Brief: Ad tech fuels disinformation sites in Europe, March 2020; https://disinformationindex.org/wp-content/uploads/2020/03/GDI_Adtech_EU.pdf.

[421] This would render it into a more structured co-regulation. ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice, May 2020; https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.

[422] European Commission, Shaping Europe's digital future: Code of Practice on Disinformation, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

[423] European Commission, Monthly Reports on Fighting COVID-19 Disinformation; https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en (Results of working with platforms) and https://digital-strategy.ec.europa.eu/en/news/coronavirus-disinformation-online-platforms-take-new-actions-and-call-more-players-join-code.

[424] The EU co-legislators politically agreed on the DMA and DSA in the trialogue in late March and April 2022 respectively. The texts of the two regulations are currently being finalized and will be formally adopted in the coming weeks. They will be directly applicable across the EU as of 6 and 15 months after entry into force. For more information, see https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

public utilities.[425] Although it is focused on illegal content and not disinformation specifically, the DSA will require illegal disinformation to being removed or its access disabled under Article 14. Moreover, the DSA requires risk assessments and appropriate mitigation measures subject to independent auditing.[426] Furthermore, the DSA bans targeted advertising on the basis of special categories of data and any targeting or amplification techniques involving the data of minors, with the aim to hamper undue interference or manipulation.[427] It does not, however, fully rule out manipulative advertising or micro-targeting entirely, or require mandatory opt-in for data-harvesting systems, which could be even more effective tools to address disinformation. At the end of the negotiations, moreover, a provision was introduced in the DSA for special measures in times of crisis such as public security or health, in which the Commission may require very large platforms "to limit any urgent threats" (for up to three months).[428] This provision demonstrates the importance and potential for manipulation or harm that the EU ascribes to platforms, both for individual opinions and public discourse, as well as a shift in the willingness for state-led interference with content governance.

The Digital Markets Act, on the other hand, aims at fair competition, and defines "gatekeepers" as core infrastructure. However, it does not include regulations on neutrality and non-discrimination, or on a fair marketplace of information or ideas.[429]

Another European legislative initiative which is currently being negotiated, the AI Act, would introduce a risk-based approach for automated tools, subjecting certain AI systems to human rights impact assessments and mitigation measures, building on product liability principles. Human rights advocates have identified the lack of individual redress or access to remedy for harms caused by AI systems as shortcomings of the current negotiation text. One proposal to address this is to add the values enshrined in Article 2 of the Treaty of the European Union (TEU) in the AI Act to enable equality bodies, ombudspersons, and national human rights institutions to be integrated into its governance system.[430] The proposal by the European Commission is currently being discussed by the co-legislators, the European Parliament and the Council of the EU, which is expected to take at least until late 2022.

---

[425] European Parliament, Disinformation and propaganda: 2021 update, p 41.

[426] Article 26 and 27 would obligate intermediaries to identify and mitigate systemic risks, including disinformation, thus leaning towards a duty of care. The DSA would also introduce transparency provisions (Article 23), relating to content moderation (Article 15) and recommender systems (Article 29), and advertising (Article 24 and 30, 31), and independent audits (Article 28). It would also facilitate supervision and research in relation to disinformation (Articles 31) and provide for redress (Article 17). Moreover, Article 35 refers to code of conducts, which may be particularly relevant in the context of disinformation.

[427] For more information, see https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users.

[428] For more information, see https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment.

[429] European Parliament, Disinformation and propaganda: 2021 update, p 40.

[430] Center for Democracy & Technology, Feedback to the EU Commission proposal on Artificial Intelligence ("AI Act"), F2665242, August 2021; https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665242_en. EDRi, Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal, January 2021; https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal. Civil society calls for integrating a rights- instead of risk-based approach to appropriately protect human rights.

The following two chapters will identify responses by states and internet intermediaries to COVID-19 online disinformation and assess them from a human rights perspective. The chapters will identify human rights-friendly approaches to addressing the challenges arising from disinformation without risking undue power, arbitrating the truth or illegitimately chilling speech.

5.2. Responses to COVID-19 disinformation by states

In order to address the rise of online disinformation related to the pandemic, states adopted regulatory and policy responses ranging from proactively providing reliable information, efforts to enhance media pluralism and digital literacy, legislation punishing the distribution of falsities, to radical disruptions of the internet.[431] Initial governmental measures included creating dedicated task forces or special units as well as outright criminalizing COVID-19 falsehoods. Some responses have sought to correct disinformation, whilst others have targeted the creators or spreaders or distribution techniques, and others focused on the audience's resilience. While some states have focused on disseminating and increasing access to evidence-based information, others restricted exactly such access or even disseminated misleading information themselves.

In general, specific responses to disinformation need to be seen in the broader context of containment measures, such as curfews, restricted access to press conferences or government information, and prolonged deadlines for information requests, which have also impacted the free flow of information, including on the pandemic.[432] Disproportionate restrictions of COVID-19-related information may be counterproductive to public health, hampering free media and restricting the public's right to receive information.[433]

5.2.1 State-led disinformation

While numerous states across the globe initiated informational campaigns, some states did not provide accurate guidance but engaged in spreading false or misleading narratives themselves. Disinformation sponsored by states and spread by sources of authority can be particularly effective, especially as states can make use of their powers, means, and reach, and combine efforts with suppressing independent sources and investigative information.[434]

In an attempt to appear more successful in the fight against the pandemic, various state actors have spread false information about their national infection rates, status of health care or fatality statistics.[435] In addition, some states disseminated unverified

[431] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 46.
[432] For analysis, see UNESCO, The Right to Information in Times of Crisis: Access to Information – Saving Lives, Building Trust, Bringing Hope!, p 9ff. Council of Europe, Press freedom must not be undermined by measures to counter disinformation about COVID-19, April 2020; https://www.coe.int/en/web/commissioner/-/press-freedom-must-not-be-undermined-by-measures-to-counter-disinformation-about-covid-19.
[433] Council of Europe, Press freedom must not be undermined by measures to counter disinformation about COVID-19.
[434] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 47ff.
[435] A/HRC/44/49, para 45

information about the origin or contagiousness of the virus or claimed refuted health advice, which may have contradicted efforts to counter the pandemic, and also undermined trust in public information and institutions.[436]

While such state-led disinformation regularly targets a state's own population, China, the Russian Federation, and Iran were accused by the European External Action Service and the US State Department of sponsoring false narratives abroad.[437] Only weeks after the novel coronavirus was first identified, the Chinese government sought to control narratives about the outbreak and spread of the virus. In an attempt to dictate an official story, the authorities allegedly withheld information, stage-managed news reports, censored informative messages, harassed journalists, detained whistleblowers,[438] and performed discrediting campaigns to shift blame away by spreading conflicting theories and amplifying existing conspiracies.[439]

Both China and Russia were also accused of fabricating reports to sow distrust, create panic and undermine confidence in the unity of the EU, while displaying themselves as "saviors" by providing medical supplies or vaccines.[440] Russia is further said to having misguided the public through selective reporting regarding the effectiveness and side effects of vaccinations by extensively reporting on the side effects of Western vaccines but not the home-made Sputnik V jab.[441] While this was aimed at portraying the Russian vaccine as superior, it may have led to an overall distrust in vaccination in the population.[442]

State-led disinformation is a clear violation of the right to freedom of expression as enshrined in Article 19 ICCPR.[443] State-sponsored foreign disinformation campaigns can additionally violate the sovereignty of another state and constitute a violation of the prohibition of intervention, or even amount to a use of force.[444] In any case, it creates an environment of fear to communicate, investigate and report, and of uncertainty over facts, thereby seriously undermining public health.[445]

---

[436] *Ibid*, para 45

[437] European Parliament, Disinformation and propaganda: 2021 update, p 30, see also table on p 34 regarding methods and funding.

[438] ARTICLE 19, Viral Lies: Misinformation and the Coronavirus, Policy Brief, March 2020; https://www.article19.org/wp-content/uploads/2020/03/Coronavirus-briefing.pdf, p 5.

[439] European Parliament, Disinformation and propaganda: 2021 update, p 29.

[440] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 371 and European Parliament, Disinformation and propaganda: 2021 update, p 28f. EUvsDisinfo, EEAS Special Report Update: Short Assessment of Narratives and Disinformation Around the COVID-19 Pandemic, April 2021; https://euvsdisinfo.eu/eeas-special-report-update-short-assessment-of-narratives-and-disinformation-around-the-covid-19-pandemic-update-december-2020-april-2021.

[441] European Parliament, Disinformation and propaganda: 2021 update, p 30.

[442] EUvsDisinfo, Kremlin Disinformation Impedes Russian Vaccination Efforts, March 2021; https://euvsdisinfo.eu/attacking-the-west-putting-russians-in-danger.

[443] Marko Milanovic, Michael N. Schmitt, Cyber Attacks and Cyber (Mis)information Operations During a Pandemic, Journal of National Security Law & Policy, Vol. 11, p. 247, May 2020; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3612019. UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 47.

[444] Milanovic et al., Cyber Attacks and Cyber (Mis)information Operations During a Pandemic.

[445] Jennifer L. Pomeranz, Aaron R. Schwid, Governmental actions to address COVID-19 misinformation, Journal of Public Health Policy, Vol. 42, Issue 2, pp 201-210, January 2021; https://doi.org/10.1057/s41271-020-00270-x, p 207.

5.2.2 Internet shutdowns

In an attempt or claim to prevent the dissemination of disinformation or incitement of violence, states have repeatedly shut down the internet entirely. Internet shutdowns, however, severely interfere with freedom of expression and media freedom, especially if aimed at silencing dissent or stifling protests.[446] Depriving access to all kinds of information and services online, including those of public interest, related to fact-finding or denouncing conspiracies, is not in line with human right standards, but inherently disproportionate.[447] While the blocking of websites or parts of services also constitute extreme measures, they may be justified under extremely narrow circumstances if no less intrusive alternative measure is available and due process is guaranteed.[448]

5.2.3 Regulatory responses

Various states enacted legislation to counteract the creation and dissemination of false information. Also a number of existing laws entail elements of prohibitions of falsehood, such as laws on consumer protection, financial fraud, defamation or perjury.[449] In recent years, several states introduced laws specifically addressing disinformation, mainly in the context of elections[450] and advertising, or regarding liability for the media or intermediaries.[451]

The German Network Enforcement Act (NetzDG) and French Loi no. 2018-2102, for example, include regulation regarding illegal hate speech and disinformation online, as does the Austrian Kommunikationsplattformengesetz (Communication Platforms Act). The U.S. Honest Ads Act or the Californian Bot Disclosure Act also address specific aspects of online deception.[452] Prior to the pandemic, about 28 states had disinformation laws in place,[453] and more than 50 countries had specific policies connected to false information.[454]

---

[446] Human Rights Council, Thirty-fifth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 30 March 2017, A/HRC/35/22; https://www.undocs.org/A/HRC/35/22, para 8ff. See also, Access Now, #KeepItO: Fighting internet shutdowns around the world; https://www.accessnow.org/keepiton. Joint Statement COVID-19: Governments must promote and protect access to and free flow of information during pandemic, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, March 2020; https://www.osce.org/representative-on-freedom-of-media/448849.

[447] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 85.

[448] Joint Declaration on Freedom of Expression and "Fake News", para 1(f).

[449] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 11f.

[450] Such as France's 2018 law on the fight against the manipulation of information in the context of elections, also Malta, Austria, Lithuania, Greece and Poland in election context. For further assessment, see Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, pp 18ff.

[451] More assessment: Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 19ff.

[452] For more information see, for example, Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 20.

[453] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression.

[454] Daniel Funke, Daniela Flamini, A guide anti-misinformation actions around the world, Poynter, August 2019; https://www.poynter.org/ifcn/anti-misinformation-actions.

The trend of regulatory responses to disinformation accelerated significantly during the COVID-19 pandemic, with at least 17 states introducing new pandemic-related laws against falsehood.[455] With several laws hastily passed, some were repealed or revised only weeks after their adoption, while others remain in force to this day.[456]

Most legislative initiatives addressing COVID-19 disinformation focus on censoring false content, with some aimed at the initial creation and others on the dissemination. Yet, these laws are regularly based on overbroad definitions. Speech restrictions on vague scopes are typically neither legitimate nor proportionate. In fact, they can facilitate arbitrary decision-making and political misuse to stifle criticism and media freedom, and strategic lawsuits against public participation (SLAPPs).[457] Disinformation laws are thus particularly concerning if adopted in a context where "fake news" claims target independent journalism and erode trust in the media's watchdog function.[458] This is illustrated by the number of journalists imprisoned on charges of "fake news", which has increased significantly in an attempt to control messaging about COVID-19.[459]

Criminalizing COVID-19 disinformation thus raises serious concerns about proportionality. Criminal law risks restraining public discussions and reporting.[460] Due to its chilling effects, criminal speech restrictions require the highest levels of justification with clear causal and culpability requirements. For the assessment of proportionality, it is relevant to evaluate free speech safeguards such as requiring malicious intent or a threshold of harm, as well as criteria to absolve liability, the burden of proof and assumption of innocence.[461] Legislation aimed at restricting online content for "creating fear", especially if this is defined by government agencies themselves, is thus eminently problematic.[462]

While legislation focusing on sharing false information may aim to minimize exposure to falsity and hence its impact, it is highly intrusive to the free flow of information. A general prohibition on disseminating information based solely on falsehood is thus incompatible with Article 19 of the ICCPR, as stated by the international free speech mandate holders.[463]

---

[455] International Press Institute, Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges, October 2020; https://ipi.media/rush-to-pass-fake-news-laws-during-covid-19-intensifying-global-media-freedom-challenges.

[456] For assessment of emergency laws with examples, see Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation.

[457] Pomeranz et al., Governmental actions to address COVID-19 misinformation, p 205. Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, p 2.

[458] Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, p 2.

[459] Committee to Protect Journalists (CPJ), Amid COVID-19, the prognosis for press freedom is dim: Here are 10 symptoms to track, June 2020; https://cpj.org/reports/2020/06/covid-19-here-are-10-press-freedom-symptoms-to-track.

[460] ARTICLE 19, Viral Lies, p 10f.

[461] OSCE Representative on Freedom of the Media, Report on International law and policy on disinformation in the context of freedom of the media, p 11.

[462] As, for example, in Thailand, see Access Now, Thailand: Stop Weaponizing 'COVID-19' to Censor Information "Causing Fear" and Crack Down on Media and Internet Service Provider, August 2021; https://www.accessnow.org/cms/assets/uploads/2021/08/Joint-Statement-Thailand-Regulation29-COVID19-August2021-FINAL.docx.pdf.

[463] Joint Declaration on Freedom of Expression and "Fake News", para 2(a).

Various assessments of legislation based on vague terms, such as anti-terrorism legislation[464] or laws addressing disinformation prior to the pandemic, identified risks of affecting legitimate speech and media freedom, be it intentionally or unwittingly.[465] Also several state measures to counter COVID-19 disinformation have been assessed to in fact asserting excessive control over the internet.[466] Content-based restrictions in the context of the pandemic[467] have regularly been deemed disproportionate by various international organizations, civil society and academia, failing to meet the three-part test of legality, legitimacy, and necessity and proportionality, for instance by not providing sufficiently precise definitions of the harm they seek to prevent, or of the nexus between falsehood and harm. Some definitions of disinformation were merely tautological, defining false information as "not true". Furthermore, some laws did not provide for judicial oversight, let alone scrutiny from parliament or national human rights institutions.[468] Other legislation involved disproportionate punishment, while less intrusive means would be available and arguably equally effective. In short, various COVID-19 disinformation laws may result in chilling effects, self-censorship and misuse to suppress criticism – and thus interfere with the right to freedom of expression.[469]

In case of a threat to the life of a nation, Article 4 of the ICCRP permits states to derogate from their human rights obligations "to the extent required by the exigencies of the situation". Several pieces of disinformation legislation, especially in the beginning of the pandemic, were indeed based on proclaimed states of emergency. A derogation must not, however, put the right itself in jeopardy.[470] Moreover, restrictions to freedom of expression still have to be in line with the three-part test as stated in Article 19(3).[471]

Derogations from the right to information are particularly difficult to justify, as such restrictions would need to be necessary precisely to respond to the threat identified. Yet, access to information is essential at all times, but especially during a crisis.[472] Emergency

[464] More details and examples of laws contradicting Article 19 of the ICCPR, see UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 53-55.
[465] For an assessment of EU responses and EU tools, for example, see European Parliament, Study on regulating disinformation with artificial intelligence, pp 34ff.
[466] Freedom Online Coalition (FOC), Joint Statement on Spread of Disinformation Online, November 2020; https://freedomonlinecoalition.com/wp-content/uploads/2021/05/FOC-Joint-Statement-on-Spread-of-Disinformation-Online.pdf.
[465] E.g. in Romania and Hungary, see Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 19.
[468] Council of Europe, Press freedom must not be undermined by measures to counter disinformation about COVID-19.
[469] See, for example, an assessment in in European Parliament, Disinformation and propaganda: 2021 update, p 110 regarding Belgium, Bulgaria, Cyprus, Denmark, France, Germany, Hungary, Italy, Ireland, Luxembourg, Portugal, Spain and the Netherlands, with 13 states declaring public emergencies; as well as in European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 39; Access Now, Recommendations on Fighting Misinformation During COVID-19; and Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 372 regarding Slovakia, Hungary, Russia, Azerbaijan, Romania, Armenia, and Bosnia and Herzegovina. See also UNESCO, Disinfodemic: Dissecting responses to COVID-19 disinformation, p 6.
[470] Human Rights Committee, General comment No. 34, para 21.
[471] A/HRC/44/49, para 16.
[472] UNESCO, The Right to Information in Times of Crisis: Access to Information – Saving Lives, Building Trust, Bringing Hope!, p 6, p 8 and sources therein, and p 16.

measures should not come at the expense of human rights or democratic guarantees,[473] let alone be used as excuse to constrain free speech and access to information.[474] Moreover, excessive regulatory responses often go hand in hand with a militaristic narrative and politicians framing the challenges from a security and foreign policy angle alone, around conflict, weapons, and defense. Such a portraying of disinformation as a hybrid threat risks disregarding human rights, which are at the core of sustainable peace, security and development.[475]

5.2.4 Third-party liability and requests for content takedowns

Over the course of the last decade, states increasingly introduced measures targeted at private actors, in particular internet intermediaries, to help them enforce policy objectives of tackling "problematic" online speech. Such measures include both legislative approaches and increased political pressure. This trend was accelerated during the COVID-19 pandemic, including in the U.S., where authorities are traditionally more reluctant to impose regulations on private actors. U.S. President Joe Biden even claimed Facebook is "killing people" by allowing COVID-19 disinformation to spread and networked anti-vaccine movements to leverage the platform's advertising and recommender systems.[476]

In order to address illegal or otherwise harmful online content, states regularly mandate internet intermediaries to take such content down.[477] Takedown requests for specific pieces of content can be based on legal provisions, systems of trusted flaggers, or be exerted through informal pressure or backdoor deals. Individual takedown requests regularly lack transparency, despite efforts to shed light on them in transparency reports and publicly available databases.[478] In principle, informal cooperation with law enforcement authorities risks circumventing the rule of law and democratic deficits.[479] The Manila Principles of Intermediary Liability, for example, do thus call for rigorous transparency for public-private partnerships while acknowledging that they can be useful in fighting problematic speech such as disinformation.[480]

---

[473] Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, p 3.
[474] For specific case studies, see Freedom House, Information Isolation: Censoring the COVID-19 Outbreak, March 2021; https://freedomhouse.org/report/report-sub-page/2020/information-isolation-censoring-covid-19-outbreak.
[475] Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 21f; EUROPOL, for example identifies disinformation as hybrid threats and part of cyberattacks, see EUROPOL, Catching the Virus: Cybercrime, Disinformation and the COVID-19 Pandemic, April 2020; https://www.europol.europa.eu/publications-documents/catching-virus-cybercrime-disinformation-and-covid-19-pandemic. See also Council of Europe, Parliamentary Assembly, Resolution 2217, April 2018; https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=24762&lang=en.
[476] Jacob Silverman, Facebook is Designed to Spread Covid Misinformation, The Soapbox, 19 July 2021; https://newrepublic.com/article/163002/facebook-designed-spread-covid-misinformation.
[477] UN Special Rapporteur on freedom of expression, A/HRC/32/38, para 85.
[478] Berkman Klein Center for Internet & Society at Harvard University, Lumen Project, https://www.lumendatabase.org.
[479] Elkin-Koren et al, Separation of Functions for AI, p 49.
[480] ARTICLE 19, Viral Lies, p 15. Manila Principles, May 2015; https://manilaprinciples.org. Public-private partnerships are particularly problematic if is they involve law enforcement regardless of whether content is illegal, see Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 20.

In addition to individual takedown requests, several states compel intermediaries to autonomously remove content they deem illegal, sometimes accompanied with threats of heavy fines should they fail to comply. By imposing third-party liability on intermediaries, states shift the legal analysis of content to private actors, away from independent courts.[481] Such outsourcing of judicial responsibilities provides intermediaries with substantial discretion and enforcement authority to govern online speech.[482] As this includes a delegation of human rights protection without due process,[483] it risks resulting in an over-removal of legitimate speech as it incentivizes intermediaries to err on the side of caution for fear of being sanctioned.[484] The prospect of future stricter regulation may additionally incentivize excessive takedowns.[485]

The human rights concerns of state-imposed takedowns is further aggravated by the fact that intermediaries tend to base takedowns rather on their terms of services than on national law, which are regularly more restrictive than what is legally required and enables the bypassing of legal constraints and oversight meant to safeguard against censorship.[486]

Moreover, such regulations often involve very short timeframes for intermediaries to remove content, which incentivizes, if not necessitates, the use of automated tools, potentially even ex ante monitoring and filtering. Mandating automation for speech governance, however, is highly problematic due to automation's lack of reliability and potential discriminatory impact.[487] Such legislation thus exacerbates existing risks to free speech and user privacy, while expanding platform authority with little oversight and countervailing checks.[488]

Generally, regulatory responses often fail to include meaningful checks while providing internet intermediaries with a source of influence and revenue.[489] Public-private cooperation and legislation to moderate content based on proprietary technology generally entrench even more power to intermediaries, rendering them indispensable actors for governments. From a human rights perspective, however, states should refrain from entrusting intermediaries with more powers without public oversight.[490] On the contrary, if any decisions over speech restrictions are mandated to intermediaries, states

[481] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 90. For an assessments of regulatory approaches in Germany, France, Italy and Spain, see European Parliament, Disinformation and propaganda, p 100.
[482] Bloch, Automation in Moderation, p 45.
[483] Haas, Freedom of the Media and Artificial Intelligence, p 4.
[484] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 58.
[485] Sanders, Human Rights-Based Approach, p 952. European Parliament, Study on regulating disinformation with artificial intelligence, p 18, Bloch, Automation in Moderation, p 61.
[486] *Ibid*, p 95.
[487] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 2; RFoM, SAIFE Policy Manual, p 1; and Gorwa et al., Algorithmic content moderation, p 2.
[488] Bloch, Automation in Moderation, p 43.
[489] *Ibid*, p 46f.
[490] Human Rights Council, Forty-first session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "surveillance and human rights", David Kaye, A/HRC/41/35, May 2019; https://undocs.org/A/HRC/41/35. See also UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348.

should ensure safeguards and clearly delineate limits for delegated decision-making power.[491]

However, despite the expected human rights infringements, evaluations of existing legislation requiring intermediaries to independently remove certain content, such as the German NetzDG, indicated less severe impacts than feared by many.[492] Other legislation that focus not only on illegal content but also introduce a general risk-based approach, such as the UK's Online Safety Bill, for example, establish a more universal duty of care for platform operators.[493] Such legal approaches may encompass responses to disinformation beyond what is illegal, also addressing the human rights risks of the algorithmic architecture and design of services.[494]

Any regulation mandating intermediaries to take down content inevitably contributes to opaque takedowns outside rule of law safeguards, resulting in a lack of data for research on how online information flows affect health or on the macro- and micro-level such as consequences of relying on automation to moderate complex content. For this reason, several civil society organizations called on intermediaries to preserve and share data.[495]

In addition to content-related restrictions, some states introduced transparency obligations for internet intermediaries regarding potentially deceptive behavior, such as obliging them to clearly label bots. Such labels aim at empowering individuals by allowing them to critically analyze the content produced or shared by bots.[496] California, for example, outlawed the non-transparent use of bots.[497] While transparency is an essential tool for user empowerment and agency, previous legislation such as the GDPR showed that making users aware does not in itself result in systematic changes.[498]

Some policymakers generally call for more traceability online, including by a real-name system. Anonymity and pseudonymity, however, can protect vulnerable voices and enable their participation in public discourse. A ban would impede such participation while doing little to hinder well-funded, sophisticated disinformation.[499] Similar concerns are raised against calls to "regulate" encryption in view of addressing COVID-19 disinformation in private messaging services. Any built-in vulnerability to encryption would automatically enable malicious actors to equally access and exploit this backdoor, with detrimental consequences for journalists, human rights defenders and individuals alike.[500]

---

[491] Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, p 12.
[492] Heidi Tworek, Paddy Leerssen, Transatlantic Working Group, An Analysis of Germany's NetzDG Law, April 2019; https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.
[493] For more information, see https://bills.parliament.uk/bills/3137.
[494] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression, p 334.
[495] Center for Democracy & Technology, COVID-19 Content Moderation Research, April 2020; https://cdt.org/insights/covid-19-content-moderation-research-letter.
[496] Nuñez, Disinformation Legislation and Freedom of Expression, p 793.
[497] *Ibid*, p 794.
[498] European Parliament, Disinformation and propaganda: 2021 update, p 41.
[499] François, Actors, Behaviors, Content: A Disinformation ABC, p 3.
[500] The importance of anonymity and encryption to exercise the rights to freedom of opinion and expression were recognized by the UN Special Rapporteur on freedom of expression, see Human Rights Council, Twenty-ninth

5.2.5 Human rights-friendly responses to COVID-19 online disinformation

While the previous chapter highlights the risks that state responses to online falsehood can entail for freedom of expression, not adopting any responses to disinformation, or letting intermediaries address the problem alone, would not be in line with states' positive obligations to enable freedom of expression. The threat that COVID-19 disinformation poses to health, public discourse and social cohesion requires state responses.[501] This includes the duty to respond to online falsehood exploiting intermediaries' architecture at odds with diversity and public interest,[502] as well as to guarantee effective pluralism[503] and a favorable environment for public debate.[504]

While any speech regulation inevitably comes with certain tradeoffs,[505] correcting information as least intrusive to the free flow of information has regularly been recommended as a response to disinformation,[506] in combination with increasing informational literacy and empowering individuals, as well as promoting quality journalism[507] and fact-checking.[508]

To address the spread of falsehood, international institutions, the free speech mandate holders and civil society regularly call on states to proactively disseminate reliable and trustworthy information,[509] while considering when and how information is best presented to be easy to understand.[510] Scientific information, in particular, should be provided in a simple and accessible way. In this context, using innovative approaches as enlisting "influencers"[511] and journalists,[512] or generating informative memes to reach broader audiences can be effective.[513] To effectively address falsehood, authoritative information provided by state actors should focus on the best available evidence and refrain from political connotations.[514] In view of restoring trust, relevant scientific

session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 22 May 2015, A/HRC/29/32; https://www.undocs.org/A/HRC/29/32.

[501] European Commission, Tackling COVID-19 disinformation – Getting the facts right, p 3.

[502] Silverman, Facebook is Designed to Spread Covid Misinformation.

[503] See, for example Refah Partisi (the Welfare Party) and Others v. Turkey, application no. 41340/98, 41342/98, 41343/98 and 41344/98, judgment, 13 February 2003; http://hudoc.echr.coe.int/eng?i=001-60936 , para 89, where the ECtHR states that "there can be no democracy without pluralism".

[504] European Court of Human Rights, Huseynova v. Azerbaijan, application no. 10653/10, judgment, 13 April 2017; http://hudoc.echr.coe.int/eng?i=001-172661, para 120.

[505] François, Actors, Behaviors, Content: A Disinformation ABC, p 1.

[506] Independent High level Expert Group on Fake News and Online Disinformation, p 37.

[507] European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 42ff (including examples). Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. Nuñez, Disinformation Legislation and Freedom of Expression, p 792.

[508] For example by promoting web-browser extensions, see UNICRI, Stop the virus of disinformation. The risk of malicious use of social media during COVID-19 and the technology options to fight it, p 21.

[509] Joint Declaration on Freedom of Expression and "Fake News", para 2(d). Joint Statement COVID-19: Governments must promote and protect access to and free flow of information during pandemic.

[510] ARTICLE 19, Viral Lies, p 13.

[511] European Parliament, Disinformation and propaganda, p 88.

[512] European Parliament, Disinformation and propaganda: 2021 update, p 58.

[513] Evidence-based communication should ensure that messages are accurate, but also catchy, says Butcher, in COVID-19 as a turning point in the fight against disinformation, p 8.

[514] Dornan, Scientific Disinformation In a Time of Pandemic, p 29.

uncertainties, medical unknowns or other knowledge limitations should be acknowledged transparently,[515] as well as science's susceptibility to revision through institutionalized procedures.[516]

In this context, Taiwan is regularly referred to as a good practice example. Taiwan's digital policies, building on multi-stakeholder participation, civil society input and democratic legitimacy, are regularly cited as human rights-friendly and effective.[517] Taiwan holds daily digital press conferences to disseminate authoritative information, and puts a strong emphasis on literacy,[518] working with traditional storytellers and providing information in local dialects, while putting efforts in reaching remote villages as well as refugee camps.[519] Taiwan has also introduced a "humor over rumor" approach and defined disinformation as "a virus" itself, that "can be caught by anyone" to avoid any blaming or polarization.[520]

If states deem legislative responses to disinformation necessary, it is often suggested to be more human rights-friendly to use existing legislation instead of rushing premature or imbalanced legislation. Privacy and data protection rules, for example, set limits to profiling and targeting, and impose transparency, accountability and redress, which can be important tools to address online disinformation.[521] Non-discrimination and antitrust regulations can also be useful in preventing targeted campaigns and undue concentration of power that may fuel disinformation.[522]

If should states instead aim to adopt new legislation targeted at COVID-19 disinformation, human rights require clear, precise and narrow definitions in order not to chill legitimate speech or encourage private actors to police speech in a way harmful to freedom of expression. In light of falsity's inherent ambiguity, definitions are difficult. From a human rights perspective, legislation should thus rather focus on the process of creating and/or disseminating disinformation than the content itself. Inspiration could be drawn from the Rabat Plan of Action to determine the severity and necessity to criminalize incitement to hate speech. Criteria such as context, the status of the speaker, their intent, the specific piece of content and form of the speech, as well as its reach, and likelihood of risk could also provide guidance for regulation aimed at limiting the creation and spread of disinformation.

With regards to third-party liability for false content, moreover, the international free speech mandate holders emphasized that they should only be imposed if the intermediary intervened on the content item or refused to obey an order from an

---

[515] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 385. The sharing of not yet peer-reviewed scientific papers can also be used for disinformation or dismantling trust if the papers do not found on strong evidence, see p 382.
[516] Dornan, Scientific Disinformation In a Time of Pandemic, p 12.
[517] For more examples and details, see European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 40f (not COVID-19 specific).
[518] Pomeranz et al., Governmental actions to address COVID-19 misinformation, p 205.
[519] *Ibid*, p 208.
[520] European Parliament, Disinformation and propaganda: 2021 update, p 54ff.
[521] Alex Campbell, How Data Privacy Laws Can Fight Fake News, Just Security, August 2019; https://www.justsecurity.org/65795/how-data-privacy-laws-can-fight-fake-news. Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 12.
[522] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 12.

independent due process oversight body, such as a court.[523] Human rights demands for limits to third-party liability have also been recognized by the ECtHR.[524] In general, broad liability rules and pressure on intermediaries risks chilling speech that could undermine health responses. A generic duty of care for intermediaries, on the other hand, could create incentives to promote user safety and mitigating social risks.[525]

In principle, while online disinformation is typically discussed in the context of content moderation, considering regulation merely through the content lens risks being incomplete and further eroding protections to freedom of expression.[526] As the current digital ecosystem risks jeopardizing exposure to a diversity of reliable voices, inspiration for responses can be drawn from previously novel technologies, such as the radio, television or satellite. The emergence of these technologies resulted in regulation to increase plurality, for example by establishing media with a public service mandate[527] and legislation on impartiality, fairness and accuracy. Some novel policy initiatives are headed in the same direction, such as the German Medienstaatsvertrag which includes a prohibition of discrimination against "journalistic editorial content".[528]

Comparable plurality regulation aiming to counteract disinformation could entail requirements on a transparent selection of content, or even a prioritization of trusted sources and public interest content.[529] While such approaches are still under-theorized, they are partly considered in the EU Code of Practice and the work by the Council of Europe.[530] Intermediaries, for instance, could be legally required to "earn" liability exemptions by proof of serving the public interest.[531] Besides, regulation to address disinformation could focus on verifications for websites with credibility scores inspired by journalistic standards.[532] However, any proactive content prioritization – be it through prominence (location of content)[533] or easier discoverability (likelihood of

[523] Joint Declaration on Freedom of Expression and "Fake News", para 1(d).
[524] In Delfi AS v. Estonia, the ECtHR states liability is in certain cases not violating Art 10, if it concerns unlawful hate speech with exceptionally strong interest in regulation, which may allow a constantly review and erasing of such comments. See European Court of Human Rights, Delfi AS v. Estonia, judgment, 16 June 2015; http://hudoc.echr.coe.int/eng?i=001-155105. In Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, however, the ECtHR notes that a news platform being compelled to police user comments in search of defamatory ones would have "chilling effect on freedom of expression on the internet", as it would incentivize over-removal, and hence infringes Art 10. The court takes into account the availability of a notice-and-take-down-system as appropriate tool to balance the rights involved and the non-profit nature of the internet content provider.
[525] ISD, Disinformation Overdose, p 45.
[526] François, Actors, Behaviors, Content: A Disinformation ABC, p 6.
[527] European Parliament, Study on regulating disinformation with artificial intelligence, p 6.
[528] For an assessment (in German), see Gahnz et al., Breaking the News? Politische Öffentlichkeit und die Regulierung von Medienintermediären, p 13ff.
[529] Giving preference to fact-checkers or media organizations, however, has already been contested within net neutrality debates, see European Parliament, Study on regulating disinformation with artificial intelligence, p 40.
[530] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 19.
[531] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 24.
[532] UNICRI, Stop the virus of disinformation. The risk of malicious use of social media during COVID-19 and the technology options to fight it, p 19.
[533] Council of Europe, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, p 18ff. Some obligations already, e.g. in Germany by Inter-State Media Treaty regarding general interest content, or UK's Prominence Rules.

consumption)[534] – also entails certain risks that need to be considered.[535] Any such efforts should hence be based on strong transparency, independent audits, and standards of separating such decisions from commercial ones.[536]

The aim of protecting the reliability of information should not revert to excessive state control, but to a democratic protections, oversight, and transparency.[537] There should be no government control over the flow of information that individuals and societies rely on to make democratic or health decisions.[538] The UN Special Rapporteur on freedom of expression clearly states that state censorship, just as disinformation, deprives individuals of autonomy, and can cause serious harm, by design or by negligence.[539]

In order to ensure evidence-based regulation, states should enable and fund research to better understand the scale and scope of online disinformation, as well as responses by state and non-state actors. The current difficulties of researchers to access data inevitably lead to a lack of evidence for policies and regulations.[540]

Besides, several international actors identified an overall shrinking civic space and crack-down of independent media due to state responses to disinformation, and misuse of pandemic-related policies to avoid public scrutiny. A shrinking space for civil society or the media, however, is not an adequate response to disinformation. On the contrary, it risks contributing to amplifying misperceptions, increasing mistrust and fostering fear.[541] A lack of checks and balances, or control vacuum typically provided by civil society, makes individuals and societies more susceptible to disinformation.[542] States should therefore promote civil society, ensure strong whistleblower protection,[543] and enable the media's watchdog role.[544] A strong rule of law, media freedom and pluralism with lively public debates make audiences more resistant to disinformation.[545]

As reliable and pluralistic information is a proven antidote to disinformation, the UN Special Rapporteur on freedom of expression emphasizes that states should provide an enabling environment for media freedom and ensure the safety of journalists.[546] Quality journalism and strong, adequately resourced independent public service media with a clear mandate to serve society can provide important alert mechanisms, fact-checking and reliable, guiding information.[547] A crisis-stricken media struggling with

[534] Council of Europe, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, p 22ff.
[535] *Ibid*, p 40ff.
[536] *Ibid*, p 47ff.
[537] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 20f.
[538] Mark MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, Transatlantic Working Group, February 2020; http://dx.doi.org/10.2139/ssrn.3615726, p 29.
[539] UN Special Rapporteur on freedom of expression, A/HRC/44/49, para 60.
[540] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 61.
[541] *Ibid*, para 85.
[542] European Parliament, Disinformation and propaganda, p 54.
[543] Access Now, Recommendations on Fighting Misinformation During COVID-19, p 15.
[544] Joint Declaration on Freedom of Expression and "Fake News", para 5(b).
[545] European Parliament, Disinformation and propaganda, p 53.
[546] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 93.
[547] European Parliament, Disinformation and propaganda, p 136.

trust, though, cannot effectively counteract disinformation.[548] State responses should thus undertake efforts to restore public trust in the media and integrity of information.[549]

While restoring confidence in fact-based journalism empowers individuals,[550] information and digital literacy can significantly strengthen individuals' and society's capacities to exert critical thinking and resistance to disinformation.[551] Literacy alone, however, cannot solve the crisis of disinformation. Informational and digital literacy should be part of a comprehensive governance system, as the responsibility should not be put on individuals alone. Individuals regularly do not have the tools to refute disinformation, particularly in complex contexts such as the COVID-19 pandemic, or if targeted by sophisticated persuasion techniques used for issue-based advertising.

Data protection remediation for individuals also remains insufficient if there are no incentives for grievances due to high personal costs and a general limited impact of individual decisions. Robust public enforcement of data protection, however, can mitigate some risks of automated content governance, and in preventing manipulation, as recognized by the EU.[552] The European Data Protection Board recognizes that data protection inherently aims at shielding against exploitation and manipulation stemming from chilled expression due to today's constant digital surveillance.[553] That said, data protection laws often only protect personal data, while automated tools also use anonymized or generated data. Data protection has also proven insufficient for systemic dark patterns or manipulation techniques to extort users' consent.[554] As the excessive use of data (personal or not) in today's data-driven economy have costs not only for individuals but entail a societal, public impact,[555] individuals' consent or redress alone cannot solve all human rights concerns.

Therefore, while facilitating collective remedy mechanisms could provide for additional redress,[556] the burden to address disinformation should not be put on individuals alone, in particular as long as market forces favor falsity. For individuals to

---

[548] *Ibid*, p 79. On the contrary, a struggling media risks unleashing all sorts of easily accessible information regardless of verification or quality, see Ghani et al., Disorder in the newsroom: the media's perceptions and response to the infodemic. It is important, however, to recognize that problematic journalism with errors arising from such sloppy verification or poor research is not the same as disinformation, even is ethics or professional standards are lacking. See, UNESCO, Journalism, 'Fake News' & Disinformation, p 8.

[549] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 86.

[550] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 94.

[551] European Parliament, Disinformation and propaganda: 2021 update, p 105f.

[552] European Commission, Communication from the Commission to the European Parliament and the Council, Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition – two years of application of the General Data Protection Regulation, June 2020; COM(2020) 264 final; https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0264&from=EN.

[553] European Data Protection Board, Opinion 3/2018, Opinion on online manipulation and personal data, p 9.

[554] In the recent report A/HRC/48/31, the UN High Commissioner for Human Rights recognizes that the use of AI poses serious threats to privacy and other human rights even when no personal data are involved. See Human Rights Council, Forty-eight session, Report of the United Nations High Commissioner for Human Rights on "The right to privacy in the digital age", Michelle Bachelet, A/HRC/48/31, September 2021; https://www.ohchr.org/EN/Issues/DigitalAge/Pages/cfi-digital-age.aspx.

[555] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 151.

[556] While this option exists in the GDPR, most EU Member States have opted against enshrining such collective remedies in line with Art 80/2 of the GDPR. See, European Parliament, Disinformation and propaganda, p 76.

be empowered, transparency, audits and meaningful oversight needs to be guaranteed.[557] Tackling COVID-19 disinformation requires strong democratic checks and balances and good governance to address undue power, and any abuse of power. Consequently, states need to realign private incentives that have transformed the digital ecosystem into a something comparable with a heavily monitored shopping mall with the public interest of a modern public agora. In short, states need to address the sociotechnical context of disinformation.

5.2.6 Addressing the sociotechnical context of COVID-19 disinformation

In an attempt to address the commercialization of the information ecosystem and commodification of personal data, often referred to as "surveillance capitalism",[558] various states and international actors introduced legislation aimed at increasing transparency in advertising and sponsored content. Misleading advertising, for example, is prohibited in various contexts.[559] Subliminal advertising is regularly outlawed, and subliminal manipulation through automated systems is proposed to be generally banned in the EU AI Act.[560] High degrees of transparency are typically required for political advertising, such as establishing publicly available ad repositories or libraries to enable public scrutiny, which may constrain certain sponsored disinformation and propaganda campaigns.[561] In November 2021, the European Commission presented a new legislative proposal on political advertising to ban any targeting and amplification that use or infer "sensitive personal data". It aims at ensuring the source and purpose of advertising is known in view of combatting disinformation and interference with democratic processes such as elections.[562] While this new proposal aims at mitigating risks of deception, a clear distinction between political and other advertising might not be fully effective from the perspective of disinformation as any opaque manipulation impacts freedom of opinion and expression.

Positive effects of advertising transparency rules remain limited as long as the default settings of the online ecosystem remain the same,[563] permitting incendiary content to keep individuals engaged, creating a source of ad revenue. As disinformation can be lucrative for malicious actors, individuals, states and internet intermediaries alike,

---

[557] Lilian Edwards and Michael Veale, Slave to the Algorithm? What a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For, Duke Law & Technology Review, Vol. 16, pp 18-84, December 2017; https://scholarship.law.duke.edu/dltr/vol16/iss1/2.

[558] Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, New York: Public Affairs, 2019.

[559] See, for example, Directive 2006/114/EC concerning misleading and comparative advertising.

[560] While it is not entirely clear from the text of the proposal, based on the recitals and public statements by the European Commission, this provision is not intended to cover subliminal content governance.

[561] European Parliament, Disinformation and propaganda: 2021 update, p 39 and European Parliament, The impact of disinformation on democratic processes and human rights in the world, p 25. The NYU Ad Observatory, however, illustrates these ad archives limited value as transparency and accountability mechanism, see NYU Ad Observatory, https://adobservatory.org/missed-ads. See also Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation.

[562] European Commission, Political advertising – improving transparency; https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Transparency-of-political-advertising_en.

[563] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 68f.

meaningfully reforming the business of online advertising may be an essential element of effective responses to COVID-19 disinformation.

Good practices of checks and balances could be obtained from journalism where editorial decision-making need to be separated from commercial processes.[564] Data protection and limits to pervasive tracking can be important tools to reorient business models.[565] A step further would be if states ban micro-targeting altogether,[566] or restrict the categories of data available for advertisers,[567] limiting "lookalike micro-targeting".[568]

In the context of disinformation in particular, it is crucial to identify the lines between permissible persuasion and unacceptable manipulation such as defined by the Council of Europe as "influence that is subliminal, exploits existing vulnerabilities or cognitive biases, and/or encroaches independence and authenticity of individual decision-making".[569] While banning micro-targeting or surveillance-based advertising would in itself not resolve disinformation, it could be an important vector to limit manipulation and discrimination at scale, while also addressing other challenges in the digital ecosystem at odds with human rights, privacy, data protection, and security.[570]

Moreover, such efforts would contribute to levelling the playing field in the digital ecosystem.[571] The power concentration of a few oligopolies acting as private arbitrators of speech that set the terms for global online speech, public discourse and access to information inevitably distorts the concept of a free marketplace of ideas.[572] Moreover, market concentration typically limits innovation, comes with civic power,[573] and exacerbates, if not triggers, various human rights concerns. The contemporary power accumulation is additionally expedited by limited transparency that fails to balance the existing information asymmetry,[574] between individuals and intermediaries as well as between intermediaries and regulators. Concentrated power in the technological space, further, risks perpetuating itself as it implies monopolized access to skilled AI developers, datasets, processing powers, and funds.[575] Consequently, human rights may require

---

[564] European Parliament, Disinformation and propaganda: 2021 update, p 115.
[565] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 92.
[566] Discussions in European Parliament's Committee on Civil Liberties, Justice and Home Affairs (LIBE), which suggests the phase out of behavioral and personalized targeting. Contextual advertising, based on keywords, language or geographical locations, would remain possible and legal. See European Parliament, Opinion by the Committee on Civil Liberties, Justice and Home Affairs (LIBE), 2020/0361(COD), July 2021; https://www.europarl.europa.eu/doceo/document/LIBE-AD-692898_EN.pdf.
[567] European Data Protection Supervisor (EDPS), Opinion 1/2021 on the Proposal for a Digital Services Act, February 2021; https://edps.europa.eu/system/files/2021-02/21-02-10-opinion_on_digital_services_act_en.pdf, p 3.
[568] European Parliament, Study on regulating disinformation with artificial intelligence, p 15 - at least for political advertising
[569] Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic, para 9(c).
[570] Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising.
[571] *Ibid*, p 32.
[572] Haas, Freedom of the Media and Artificial Intelligence, p 2.
[573] European Parliament, Disinformation and propaganda, p 124.
[574] François, Actors, Behaviors, Content: A Disinformation ABC, p 7.
[575] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 8. A handful of intermediaries globally hold and control detailed personal profiles about billions of individuals and the networked public sphere, see UN High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/48/31, para 34ff.

addressing oligopolistic positions that unavoidably raise concerns for diversity of sources and views. Similar concerns were recognized in the context of legacy media, which is why European states are mandated to take appropriate action "to prevent undue media dominance or concentration".[576] Besides, concentrated control over content inevitably renders surveillance, profiling, and micro-targeting more ubiquitous and sophisticated, particularly in countries and environments where certain intermediaries' services practically constitute the entire internet or where digital authoritarianism advances and opaque public-private partnerships are prevalent.[577]

Consequence, state responses to content governance should not entrench intermediaries with more "opinion power".[578] At the same time, however, limiting intermediaries' control over the flow of information should not lead to increased state control over information or communication.[579] From a human rights perspective, regulation should rather focus on separating social infrastructure from distributing content[580] and empowering individuals, for instance through interoperability.[581] Genuine interoperability would allow for more alternatives and thus for the market forces of consumer choice to work. In this context, network and lock-in effects as well as switching costs[582] should be considered, as critical mass is needed for the profitability of intermediary services.[583] Alternatively, due to a few intermediaries' control over public discourse, they could also be regulated as essential service facilities carriers or even public utilities.[584]

5.3. Responses to COVID-19 disinformation by internet intermediaries

5.3.1 Introduction

Over the course of the last years, intermediaries increasingly departed from being mere conduits of content to hybrid actors that curate, filter and act as gatekeepers to information. This trend was further accelerated by the COVID-19 pandemic, where intermediaries gradually took on functions analogous to legacy media.[585]

---

[576] General comments No 34, para 40.
[577] Deborah Brown, Big Tech's Heavy Hand Around the Globe, Human Rights Watch, September 2020; https://www.hrw.org/news/2020/09/08/big-techs-heavy-hand-around-globe.
[578] Natali Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power, Digital Journalism, Volume 6, Issue 6, pp 842-854, July 2020; https://doi.org/10.1080/21670811.2020.1773888.
[579] The 2021 Freedom on the net report, for example, concludes that the global drive to control big tech regularly leads to censorship and surveillance and that the shift in power from companies to states has resulted in a record-breaking crackdown on freedom of expression. See Freedom House, Freedom on the net report 2021, September 2021; https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech#Internet.
[580] Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power, p 850.
[581] This is also suggested by the UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 21.
[582] Electronic Frontier Foundation (EFF), Cory Doctorow, Facebook's Secret War on Switching Costs, August 2021; https://www.eff.org/deeplinks/2021/08/facebooks-secret-war-switching-costs.
[583] Ofcom, Use of AI in Online Content Moderation, p 61f.
[584] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 21.
[585] Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, p 20.

The global pandemic may be seen not only as a turning point of intermediaries' responses to online falsehood and deceit but also of content governance in general. This fuels important questions on whether intermediaries should have to follow principles of diversity, or provide high-quality and reliable sources. In spite of increasing demands to define intermediaries as "media actors", however, legacy media's independence and autonomy should remain upheld.[586]

Following an introduction in the United Nations Guiding Principles on Business and Human Rights and intermediaries' commitment to respect human rights, this chapter will assess the responses of internet intermediaries, and social media platforms in particular, to online COVID-19 disinformation.

5.3.2 Business and Human Rights

International human rights law binds states to respect, protect, and promote individuals' rights. While states are the duty-bearers to protect human rights, companies too, have a responsibility to respect human rights as defined by the UN Guiding Principles on Business and Human Rights (UNGP). The UNGP provide minimum standards for corporate responsibility, by mandating companies to identify and mitigate actual and potential adverse human rights impacts of their operations, through human rights due diligence, communication and remediation processes. The UNGP require companies to avoid infringing human rights directly, but also to address negative impacts caused or contributed to as a result of their business activities.[587] Although the UNGP apply to all corporations, irrespective of size, turnover or revenue, they also acknowledge that mitigation measures may depend on "size, sector, operational context, ownership and structure", as well as the severity and likeliness of identified risks.[588]

The UNGP aim to direct innovation towards human rights-compliant technologies, considering human rights protection more important than individual sectoral innovations.[589]

At the same time, however, private actors enjoy the economic freedom to innovate and conduct a business, which is also to be considers when assessing potential violations of freedom of expression.[590]

5.3.3 Intermediaries' commitment to respect human rights

---

[586] *Ibid*, p 25.

[587] Principle 17-18 UNGP, see also A/HRC/17/31.

[588] Principle 14.

[589] It is often said that Silicon Valley lives by the mantra of 'move fast and break things', see European Parliament, Study on regulating disinformation with artificial intelligence, p 19. While the UNGP do not aim to stifle innovation, various sectorial restrictions exist, such as in the medical sector, for example regarding cloning. See Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Feasibility Study, CAHAI(2020)23, December 2020; https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da, para 141.

[590] European Court of Human Rights, Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary. Wolfgang Benedek, Matthias C. Kettemann, Freedom of Expression and the Internet (2nd edition), September 2020, p 156.

Already prior to the COVID-19 pandemic, all large intermediaries formally committed themselves to respect human rights. Facebook, for example, declares to take the international human rights framework into account not only to decide which content should be removed, but also to identify situations where content should remain online that otherwise would not be permitted on its platform.[591]

In addition, most intermediaries commit themselves to some sort of ethics, including in the context of the use of automation. Emerging ethical standards can provide useful internal or even industry frameworks[592] and transform ad hoc reactive measures into principled and structured approaches.[593] While such frameworks are important, their adoption has often been accused of being a tactical move to avoid strict regulation. Unless aligned with a human rights-based approach,[594] they are often considered attempts at "ethics-washing".[595]

Whistleblowers and journalists repeatedly revealed how intermediaries failed to pay sufficient attention to human rights risks caused by their services or to successfully enforce their rules against deceptive content or activities. In the context of Facebook, for example, whistleblower Sophie Zhang unveiled that the resource-intensive enforcement of inauthentic behavior rules was frequently put off where there was little PR risk, and that substantial decision-making power was shifted to individual Facebook employees without any oversight, resulting in arbitrariness.[596] Consequently, international actors, civil society, and governments continuously call on intermediaries to do more to protect human rights online.[597]

5.3.4 Specific responses to COVID-19 disinformation

On 1 April every year, Google shows an April Fools' joke. In 2020 and 2021, it did not. Instead, Google and YouTube displayed links to authoritative COVID-19

---

[591] Facebook, Updating the Values That Inform Our Community Standards, September 2019; https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standard.

[592] Council of Europe, Algorithms and Human Rights, p 41f. OSCE Representative on Freedom of the Media, #SAIFE – Spotlight on Artificial Intelligence and Freedom of Expression, p 43. Status quo means companies building constitutions for digital lands. IEEE (Institute of Electrical and Electronics Engineers) AI space standards for ethical algorithms, e.g. the FAT-ML (Fairness, Accountability, Transparency in machine learning), or IEEE 7000 scheme on algorithms, transparency, privacy, bias and ethical system design, the world's first standard showing tech companies how to build technology bringing human and social value. The IEEE is the world's biggest engineering association with over 400,000 members, see Institute of Electrical and Electronics Engineers (IEEE), 7000™-2021 Standards: Addressing Ethical Concerns During Systems Design, September 2021; https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html?utm_source=POLITICO.EU&utm_campaign=52a2ab7f3a-EMAIL_CAMPAIGN_2021_09_15_08_59&utm_medium=email&utm_term=0_10959edeb5-52a2ab7f3a-190567583.

[593] Suzor, Lawless: the Secret Rules That Govern Our Digital Lives, p 173.

[594] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, p 16.

[595] Ben Wagner; Ethics As An Escape From Regulation. From "Ethics-Washing" to Ethics-Shopping?, Amsterdam University Press, December 2019; https://doi.org/10.1515/9789048550180-016.

[596] BuzzFeed News, "I Have Blood on My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation.

[597] See, for example, EFF, Content Moderation is Broken, Let Us Count the Ways; UN Special Rapporteur on freedom of expression, A/HRC/38/35; Joint Declaration on Freedom of Expression and "Fake News"; UN Special Rapporteur on freedom of expression, A/74/486.

information.[598] Intermediaries including Facebook, LinkedIn, Microsoft, Reddit, Twitter and Google even established a coalition against COVID-19 disinformation, committing themselves to remove content that is "harmful for the health of people" and to instead provide trustworthy information.[599]

While intermediaries were long reluctant to address disinformation spreading on their services or to take on any editorial role, they took unprecedented steps to remove, hide or restrict false content since the beginning of the COVID-19 outbreak. On top of that, all big internet intermediaries cooperated closely with health authorities to promote official information and verified sources.[600] Whereas efforts against disinformation were already significantly boosted in the aftermath of the 2020 U.S. elections and the storming of the Capitol in January 2021[601], intermediaries took steps against COVID-19 disinformation with greater speed and scale than with any other false content in the past.[602] Accordingly, COVID-19 constitutes a potentially lasting turning point in intermediaries' content governance.

Already prior to the pandemic, all large intermediaries had policies in place to address disinformation,[603] but were regularly accused to fail to enforce them.[604] In an attempt to combat COVID-19 disinformation, intermediaries enforced existing rules more strictly and additionally modified algorithms to limit the virality of untruthful content.[605] Intermediaries adjusted existing policies, with many introducing new categories around COVID-19 disinformation.[606] Policy adjustments ranged from deprioritizing disinformation or adding warnings, labels or contextual information (Twitter, Facebook/Instagram, and TikTok), to removing deceptive content or accounts or reducing

---

[598] Wikipedia, List of Google April Fools' Day jokes, https://en.wikipedia.org/wiki/List_of_Google_April_Fools%27_Day_jokes#2020%E2%80%9321:_cancellation.

[599] Joint industry statement on COVID-19 from Microsoft, Facebook, Google, LinkedIn, Reddit, Twitter and YouTube, March 2020; https://twitter.com/Microsoft/status/1239703041109942272.

[600] Butcher, COVID-19 as a turning point in the fight against disinformation, p 7.

[601] These special events leading to significant policy adjustments again illustrate the strong US-centric view and focus of large internet intermediaries.

[602] UNESCO, Journalism, press freedom and COVID-19, p 4.

[603] For an analysis of "fake news" policies in intermediaries' terms of services see ARTICLE 19, Side-stepping rights, p 27ff.

[604] CCDH, The Disinformation Dozen. ARTICLE 19, Viral Lies, p 14.

[605] For an assessment of labelling as "nutrition facts"-style information for online content, and comparisons with product labelling, see Matthew Spradling, Jeremy Straub, Jay Strong, Protection from 'Fake News': The Need for Descriptive Factual Labeling for Online Content, Future Internet, Volume 13, 142, May 2021; https://doi.org/10.3390/fi13060142.

[606] Twitter, for example, differentiates between false/misleading information about the nature of the virus, about preventive measures/treatments/precautions, about official regulations, about the prevalence/risk/infection and false/misleading affiliation, see Twitter, COVID-19 policies; https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy. Facebook, for example, provides a timeline of their actions against misinformation (inter alia), see Facebook, Timeline: Taking action to combat misinformation, polarization, and dangerous organization, April 2021; https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Efforts-to-Combat-Misinformation-Polarization-and-Dangerous-Organizations.pdf. Google/YouTube differentiates between false information about the treatment, prevention, diagnosis, transmission, social distancing/self-isolation guidelines and the existence of COVID-19, see Google/YouTube, COVID-19 policies, https://support.google.com/youtube/answer/9891785?hl=en, https://support.google.com/youtube/answer/9803260?hl=en and https://support.google.com/youtube/answer/9803260?hl=en#:~:text=If%20your%20content%20receives%20a,is%20now%20eligible%20for%20monetization.

its visibility (Google, Facebook/Instagram, and TikTok), to demonetizing falsities (Google/YouTube, and TikTok), and to prioritizing authoritative information and providing FAQs.[607]

Several of these efforts build on close cooperation with fact-checking organizations and trusted flaggers.[608] Facebook, for example, heavily relies on fact-checking by independent third parties (e.g., Full Fact),[609] and continuously notes its strict stance against false COVID-19 information "that might lead to imminent harm" on its services, even following overturning decisions by the Facebook Oversight Board.[610] Adopting a slightly different approach, Twitter recently announced a new community-based approach to fight falsities (Birdwatch).[611]

In addition, many intermediaries increased their willingness to temporarily suspend or block accounts spreading disinformation. Before taking such steps, intermediaries often follow a nuanced strike system, which is regularly opaque, varies heavily among different platforms (sometimes three, sometimes five, and sometimes even more strikes are required before a suspension is initiated, for example for less severe falsities in the COVID-19 context) and can be rather easily circumvented.[612]

In an attempt to find the right balance between decreasing exposure to falsehood and excessively restricting the free flow of information in their policies, intermediaries often declare exemptions to their disinformation policies if personal anecdotes are shared, if content is based on strong opinions without false or misleading assertions of fact, or refers to debates about scientific research.[613] Intermediaries also regularly exempt

---

[607] For a detailed overview of adjusted policies based on intermediaries own reporting, see the monthly reporting by Facebook, Google, Microsoft, TikTok and Twitter based on the EU Code of Practice against Disinformation. See European Commission, Monthly Reports on Fighting COVID-19 Disinformation and internet intermediaries' COVID-19 policies.

[608] ARTICLE 19, Side-stepping rights, June 2018; https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf, p 5.

[609] Full Fact, Report on Facebook's Third-Party Fact-Checking programme, December 2020; https://fullfact.org/blog/2020/dec/full-fact-publishes-new-report-on-facebooks-third-party-fact-checking-programme.

[610] The Facebook Oversight Board overturned a decision to remove a post saying that hydroxychloroquine and azithromycin were effective COVID-19 treatments, see https://oversightboard.com/news/325131635492891-oversight-board-overturns-facebook-decision-case-2020-006-fb-fbr. The decision includes a recommendation to define key terms as "misinformation" and adopt less intrusive means of enforcements than removals, which Facebook disagrees with and stated not to take action on, see Facebook, Facebook's Response to the Oversight Board's First Set of Recommendations, February 2021; https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations. For an analysis of Facebook's responses to the Oversight Board's decisions, see Evelyn Douek, The Oversight Board Moment You Should've Been Waiting For: Facebook Responds to the First Set of Decisions, February 2021; https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-facebook-responds-first-set-decisions.

[611] Twitter Blog, Keith Coleman, Introducing Birdwatch, a community-based approach to misinformation, January 2021; https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[612] CNN, Claire Duffy, For misinformation peddlers on social media, it's three strikes and you're out. Or five. Maybe more, September 2021; https://edition.cnn.com/2021/09/01/tech/social-media-misinformation-strike-policies/index.html. See also, for example Twitter, COVID-19 policies with different consequences depending on the number of strikes and severity and type of the violation.

[613] For example Twitter, see Twitter, COVID-19 policies.

restrictions for content otherwise violating their policies on false content if part of educational, documentary or artistic purposes.[614]

Furthermore, in addition to responding to the spread of falsehood, intermediaries also increasingly cooperate with health authorities and news organizations to rank more authentic content and provide more curated content.[615] Intermediaries have elevated authoritative content,[616] for example by linking to the WHO or local health authorities by using geolocations.[617] Throughout the pandemic, almost all large intermediaries have actively promoted fact-checked news[618] and financially supported fact-checkers and journalism.[619] Google, for instance, introduced a medical topics policy; Facebook provides journalism with more prominence in its News section; Instagram provides a COVID-19 Information Center; and Mozilla ('Pocket' and 'Snippet') and TikTok provide their own informational pages and a collection of trustworthy sources.[620] Various intermediaries additionally provide their users with preventive and vetted health guidelines.[621]

In addition to that, several intermediaries introduced specific policies related to COVID-19 vaccines,[622] supporting national vaccination campaigns or providing information about vaccination locations, for example, on Google Maps.[623] Facebook provides a COVID-19 Information Center[624] with detailed and somewhat locally adapted information on the vaccine.[625]

While all these efforts of aggregating content relevant to fight COVID-19 are undeniably helpful to users, they constitute a clear deviation from intermediaries' previous hesitation to undertake editorial, media-like actions.[626] At the same time, however, in contrast to the heavily regulated media sector (around advertising, ownership, transparency and subject to accountability mechanisms),[627] intermediaries typically only face self-regulation, if at all.[628]

---

[614] See, YouTube, COVID-19 policies.
[615] European Parliament, Disinformation and propaganda: 2021 update, p 82.
[616] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 375.
[617] *Ibid*, p 375.
[618] Council of Europe, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, p 6f.
[619] European Parliament, Disinformation and propaganda: 2021 update, p 86.
[620] See Facebook, COVID-19 policies; Twitter, COVID -19 policies; Google/YouTube, COVID-19 policies; and TikTok, COVID19 policies; https://www.tiktok.com/safety/en-us/covid-19. See also Instagram, Helping People Stay Safe and Informed about COVID-19 Vaccines, March 2021; https://about.instagram.com/blog/announcements/continuing-to-keep-people-safe-and-informed-about-covid-19. For a detailed overview of policies to fight COVID-19 disinformation, see European Commission, Monthly Reports on Fighting COVID-19 Disinformation.
[621] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 375.
[622] Virality Project, COVID-19 vaccine policies.
[623] See, European Commission, Monthly Reports on Fighting COVID-19 Disinformation and Google's July 2021 submission in particular.
[624] For Facebook's COVID-19 response, including the Information Center and Facebook's plans to help get people vaccinated can be found here on https://about.fb.com/news/tag/covid-19/, see also Facebook, COVID-19 policies.
[625] Facebook, Reaching Billions of People With COVID-19 Vaccine Information, February 2021; https://about.fb.com/news/2021/02/reaching-billions-of-people-with-covid-19-vaccine-information.
[626] European Parliament, Disinformation and propaganda: 2021 update, p 86.
[627] European Parliament, Study on regulating disinformation with artificial intelligence, p 9.
[628] *Ibid*, p 14.

Additional initiatives by intermediaries to counter COVID-19 disinformation include banning specific hashtags (Instagram),[629] not autocompleting anti-vaccine hashtags (TikTok)[630] or promoting certain hashtags (Twitter) and nudges in cooperation with the WHO.[631] Intermediaries have also supported literacy programs and revised advertising policies,[632] and provided health authorities with free advertising.[633] Additionally, intermediaries strengthened efforts to stop individuals from profiting financially from COVID-19 disinformation, aimed at clickbait and counterfeit news sites for example,[634] or introducing advertising disclaimers on COVID-19 related content (Facebook/Instagram).[635]

In an attempt to address disinformation in private groups or invite-only sections that typically face little to no oversight, several intermediaries introduced policies for greater transparency regarding political advertising. Facebook, additionally, declared to display information boxes linking to official health advice in pages concerning COVID-19 content.[636]

Similarly, several measures aimed at tackling disinformation spread via end-to-end encrypted messaging services are particularly difficult to counteract due to their closed nature.[637] Such services have been exploited to conceal the real scale of the pandemic,[638] or to incite violence, such as attacks on 5G masts blamed for spreading the virus.[639]As a response, WhatsApp (Facebook) further developed its rules limiting the number of permitted forwards and labels chain messages.[640] According to a MIT study (conducted prior to the pandemic), limiting the possibility to forward messages can be effective to slow the spread of disinformation.[641] Increasing efforts have also been undertaken on

[629] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 375.
[630] TikTok, COVID-19 policies.
[631] European Parliament, Disinformation and propaganda: 2021 update, p 86. For overview of good practices in COVID-19 infodemic context by intermediaries, see European Parliament, Disinformation and propaganda: 2021 update, p 87f.
[632] UNESCO, Journalism, press freedom and COVID-19, p 5.
[633] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 375, see also Alexandre De Streel, Elise Defreyne, Hervé Jacquemin, Michèle Ledger, Alejandra Michel, Alessandra Innesti, Marion Goubet, Dawid Ustowski, Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform, European Parliament, study requested by the IMCO committee, June 2020; https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf, p 63f.
[634] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression, p 197.
[635] Facebook, COVID-19 policies.
[636] POLITICO, Mark Scott, Facebook's private groups are abuzz with coronavirus fake news, March 2020; https://www.politico.eu/article/facebook-misinformation-fake-news-coronavirus-covid19.
[637] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O. S. Vaz de Melo and Fabricio Benevenuto, Can WhatsApp Counter Misinformation by Limiting Message Fowarding?, September 2019; https://arxiv.org/pdf/1909.08740.pdf.
[638] Nithin Coca, Disinformation from China floods Taiwan's most popular messaging app, Coda Story, October 2020; https://www.codastory.com/authoritarian-tech/taiwans-messaging-app.
[639] Nazia Parveen and Jim Waterson, UK phone masts attacked amid 5G-coronavirus conspiracy theory, April 2020, The Guardian; https://www.theguardian.com/uk-news/2020/apr/04/uk-phone-masts-attacked-amid-5g-coronavirus-conspiracy-theory.
[640] Facebook, COVID-19 policies.
[641] De Freitas Mel et al, Can WhatsApp Counter Misinformation by Limiting Message Fowarding?

voice messaging apps, for example to prominently display information from credible sources.[642]

Just as intermediaries' overall content governance, so too are their responses to COVID-19 disinformation widely supported by automated systems. Since the outbreak of the pandemic, intermediaries deployed even more automation and adjusted their enforcement methods.[643] Due to curfews and social distancing rules, intermediaries relied more heavily on automation, including to detect disinformation, as well as to remove content, in particular copies of content previously flagged as disinformation.[644] While intended as temporary measure, the use of such automated tools may turn permanent once tested in a crises.[645]

It can thus be ascertained that COVID-19 fundamentally altered the way internet intermediaries govern content, as they are taking on a more active role in policing content that could have a negative effect on individuals' health, in promoting accurate and authoritative information, and in fostering a healthier online environment. These efforts, however, are limited to content related to COVID-19. Other falsehoods may continue to be promoted based on the attention economy which often values emotional resonance and controversy above public interest.[646]

At the same time, intermediaries' responses to COVID-19 disinformation have demonstrated that their gatekeeping power over the flow and accessibility of information is heavily informed by design choices and that centralized decisions can in fact determine rules for public discourse. In general, design settings can organize content based on intermediaries' private interest, to protect diversity or driven by any other interest, and does therefore clearly entail risks of undue interference if deployed without oversight and safeguards.[647]

Following intermediaries' long argument of being neutral conduits of content, this shift in content governance may result in fundamentally novel content curation approaches or lead to new state regulations requiring intermediaries to make editorial decisions also in contexts other than the COVID-19 pandemic. The first steps in this direction can already be observed, Google/YouTube for example announced at the end

---

[642] Ann Cathrin Riedel, Behind closed curtains, Disinformation on messenger services, Friedrich Naumann Foundation For Freedom, August 2020; https://www.freiheit.org/iaf/behind-closed-curtains-disinformation-messenger-services.
[643] See, for example, Twitter, An update to the Twitter Transparency Center, July 2021; https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center?utm_source=POLITICO.EU&utm_campaign=a78ce40264-EMAIL_CAMPAIGN_2021_07_15_11_42&utm_medium=email&utm_term=0_10959edeb5-a78ce40264-190567583.
[644] For example, Facebook claims to have removed more than 12 million pieces of content in that way, see Facebook, Taking Action Against People Who Repeatedly Share Misinformation, May 2021; https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation. In general, more reliance on AI, though it might lead to higher error rates and delays in appeals, see Google/YouTube, Protecting our extended workforce and the community, March 2020; https://blog.youtube/news-and-events/protecting-our-extended-workforce-and. See also Twitter, Guy Rosen on restoring incorrectly removed COVID-19 posts, March 2020; https://twitter.com/guyro/status/1240088303497400320?s=20.
[645] Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, p 3.
[646] Council of Europe, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, p 7.
[647] *Ibid*, p 10.

of September 2021 that it would remove content spreading mistruth about any vaccines' efficiency or harmfulness in the same way as COVID-19 vaccine disinformation.[648]

5.3.5 Assessing the effectiveness of intermediaries' measures

Despite intermediaries' significant efforts to tackle COVID-19 falsehood online, several studies demonstrate their failure to comprehensively do so. According to the Oxford Internet Institute less than 1% of deceptive videos have been detected and removed from Facebook.[649] In the same vein, the fact-checking organization Avaaz stated that Facebook put warning labels only on 16% of their debunked false health information,[650] in particular if cross-shared across social media.[651]

At the same time, COVID-19 information has been wrongly taken down as falsehood, both due to a lack of scientific or medical qualification of human moderators and an inadequate use of automation.[652] This resulted in high error rates particularly in less spoken languages. Avaaz found that COVID-19 disinformation is twice as likely to stay on Facebook in the EU as in the U.S.,[653] and no studies are available for even less prioritized contexts.

The Facebook Files investigations recently published by the Wall Street Journal and particularly whistleblower Frances Haugen exposed how the company is made aware of a variety of flaws causing harm, but ignores or fails to address them. This includes, for example, Facebook's aim of promoting COVID-19 vaccination while studies show how anti-vaccine activists incite vaccine hesitancy and sow doubts by exploiting Facebook's dynamics for virality.[654]

A proper assessment of intermediaries' responses to COVID-19 disinformation requires the evaluation of their use of automation. When the enforcement of policies

---

[648] Google/YouTube, Managing harmful vaccine content on YouTube, September 2021; https://blog.youtube/news-and-events/managing-harmful-vaccine-content-youtube/?utm_source=POLITICO.EU&utm_campaign=222ef59fd6-EMAIL_CAMPAIGN_2021_09_30_11_37&utm_medium=email&utm_term=0_10959edeb5-222ef59fd6-190567583.

[649] Aleksi Knuutila, Aliaksandr Herasimenka, Hubert Au, Jonathan Bright, Philip N. Howard, COVID-Related Misinformation on YouTube: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies, Oxford Internet Institute, September 2020; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/09/Knuutila-YouTube-misinfo-memo-v1.pdf.

[650] Avaaz, Facebook's Algorithm: A Major Threat to Public Health.

[651] For example, in February 2021, an integrity worker presumably flagged a post as false with 53,000 shares and three million views that was missed by automated moderation tools, see Gizmodo, Tom McKay, Internal Facebook Documents Show How Badly It Fumbled the Fight Against Anti-Vaxxers: ReportSeptember 2021; https://gizmodo.com/internal-facebook-documents-show-how-badly-it-fumbled-t-1847696500.

[652] UK House of Lords, Communications and Digital Committee, Free for all? Freedom of expression in the digital age, July 2021; https://committees.parliament.uk/publications/6878/documents/72529/default, p 9 (YouTube) and p 16 (Facebook).

[653] Avaaz, Left Behind: How Facebook is neglecting Europe's infodemic, April 2021; https://secure.avaaz.org/campaign/en/facebook_neglect_europe_infodemic: 31% in Italian, 42% in French and 67% in Spanish, almost as good as in English with 71% (even here difference whether in US or British or Irish Facebook users). See also Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation, p 12.

[654] The Wall Street Journal, The Facebook Files, September 2021; https://www.wsj.com/articles/the-facebook-files-11631713039. According to various news report, whistleblowers Frances Haugen has been invited to the U.S. Congress and EU Parliament and urged policymakers to regulate the social media giant, see POLITICO, Clothilde Goujard, Facebook faces wrath of EU lawmakers working on online content rules, October 2021; https://www.politico.eu/article/facebook-faces-wrath-of-eu-lawmakers-working-on-online-content-rules.

such as flagging, filtering, blocking, and (de)prioritizing disinformation is automation-driven, an additional layer of risks to human rights arises. Such risks are even greater if automated tools are deployed for complex tasks as fact-checking,[655] and if human involvement and review is limited. [656]

To this day, automated systems remain limited in their capacity to analyze content and have proven incapable to accomplish on a large scale the sophisticated, nuanced speech analysis that humans can perform on a small scale.[657] Therefore, automation may be particularly unfit to govern scientific information and uncertainties surrounding an ongoing health crisis with constantly evolving knowledge.[658] Consequently, outsourcing to automation decisions to distinguish between malicious and truthful information or between scientific disagreement and falsehood has serious human rights implications. This was also acknowledged by intermediaries themselves, who, in the beginning of the pandemic, announced that the increased reliance on automation would result in higher errors rates.[659] Although certain margins for error in content governance seem unavoidable,[660] high error rates risk undermining the credibility and legitimacy of responses to disinformation.

Despite the quasi-global application of intermediaries' terms of services, rules "banning" false information or deceptive practices are not enforced consistently across all jurisdictions and geographic areas.[661] On the other hand, a global, undifferentiated and automated enforcement of policies without sufficient adjustments to local contexts risks disproportionally affecting less prioritized communities and particularly marginalized groups due to the incapacities of automated decision-making systems to understand nuances.[662]

Several intermediaries have undertaken internal assessment processes regarding their policies. Facebook, for example, presented internal findings that providing "related articles" next to debunked news stories prove to result in fewer shares than simply showing a disputed flag.[663] Twitter, as another example, announced that its new requirements to accessing a link or typing one's own text before retweeting posts significantly decreased the spread of disinformation. While this illustrates how interface changes can be effective responses to online falsehood, such assessments are regularly done by intermediaries themselves without transparency to the outside. Such design

[655] European Parliament, Study on regulating disinformation with artificial intelligence, pp 42ff.
[656] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 71.
[657] CDT, Mixed Messages? The Limits of Automated Social Media Content Analysis, p 3.
[658] Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty, p 383. For example, statements referring to the potential origin of the virus in an virology institute have been removed or marked as falsity until April 2021, when the scientific discussion shifted. See UK House of Lords, Communications and Digital Committee, Free for all? Freedom of expression in the digital age, p 17.
[659] See UNESCO, Disinfodemic: Deciphering COVID-19 disinformation, p 11 and sources therein (announcements by intermediaries).
[660] Mike Masnick, Impossibility Theorem; Content Moderation at Scale is Impossible To Do Well, Tech Dirt, November 2019; https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml.
[661] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 65 and 70.
[662] Caplan, Content or Context Moderation?, p 25.
[663] European Parliament, Disinformation and propaganda, p 104 and sources therein.

changes are seldom if ever announced or declared, which means users are somehow treated as guinea pigs for policy experiments.[664]

Additional challenges to independently assess intermediaries' responses to COVID-19 disinformation arise from the fact that policies are updated almost on a weekly or even daily basis, sometimes only in specific regional contexts. These constant changes make rules unpredictable, especially as they do not tend to be disclosed[665] and information on policies remains hard to find.[666]

The effectiveness of responses to COVID-19 disinformation also remains difficult to assess as measures such as providing reliable information are part of broader efforts to increase people's awareness of health measures. Additional difficulties stem from the fact that definitions, and hence responses, vary across intermediaries' services.[667]

An independent assessment is further impaired as disinformation rules are not always consistently applied to all users, including public figures. While international human rights law provides for specific protection of political speech and issues of public interest, such decisions are regularly neither clearly set out nor disclosed.[668]

The current power asymmetries of the digital ecosystem, with a few dominant intermediaries governed by a handful of white male millionaires in Silicon Valley with little internal checks and balances, let alone independent oversight, is further unconducive to independent assessment.

Despite general challenges to assess responses to COVID-19 disinformation, the UN Special Rapporteur on freedom of expression determined that reactive content moderation alone proved to be insufficient.[669] Given that intermediaries' business models underpin much of the drivers of disinformation, effective responses need to go beyond removing or down-ranking false information online. In particular, moderation tools such as removals on the front-end are regularly unsuccessful as content may continue to be spread and be lucrative through the back-end commercial and technological infrastructures.[670]

Accordingly, more effective content moderation remains a downstream effort that ultimately falls short of solving the information disorder as long as upstream systems are designed for automated amplification and audience targeting.[671] From a human rights perspective, it is thus necessary to review the entire infrastructural facilitation of

---

[664] François, Brookings Podcast on COVID-19 and the ABCs of disinformation.

[665] For an overview and timelines or changes transparently reported (which do not cover all changes), see European Commission, Monthly Reports on Fighting COVID-19 Disinformation. See also Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 155f.

[666] This is, for example, a repeated criticism in the Ranking Digital Ranking Accountability Index. See 2020 Accountability Index, April 2021; https://rankingdigitalrights.org/index2020.

[667] Facebook, for example, rather focuses on addressing inauthentic behavior than false information. See Facebook, COVID-19 policies.

[668] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 77, 79. The Wall Street Journal, The Facebook Files.

[669] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 65.

[670] For more information on lucrative online disinformation methods, see chapter 3.3, Au et al, Profiting from the Pandemic, and Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 16.

[671] Nathalie Maréchal, Rebecca MacKinnon and Jessica Dheere, Getting to the Source of Infodemics: It's the Business Model, Ranking Digital Rights, New America, May 2020; https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model.

disinformation;[672] responses solely addressing the most visible aspect of disinformation remain insufficient.[673]

5.3.6 General human rights safeguards for content governance

In order to ensure human rights-compliant speech restrictions for COVID-19 disinformation, intermediaries could apply the criteria of legality, legitimacy and necessity and proportionality in line with Article 19 of the ICCPR. This would entail that content governance are based on precise and accessible rules that provide information on the different categories of disinformation and the possible reactions, ideally with examples and detailed guidance.[674] The UN Special Rapporteur on freedom of expression underlined that also private speech restrictions should pursue legitimate aims, such as the rights and reputation of others, upholding pluralistic public discourse and guaranteeing non-discrimination.[675] This would mean that content governance rules as well as individual restrictions need be proportionate, using the least intrusive measures, i.e., go beyond binary decisions of removals or blocking towards a more context-sensitive balancing and take into account the pandemic-specific situation and potential harm of disinformation.

Although entrepreneurial freedoms to some extent allow maintaining a business according to own paradigms and to build a community as desired, media regulators emphasized that large internet intermediaries, whose policies influence global public debate, have some societal responsibility and thus have to set higher standards of public interest.[676] In the same vein, the concept of "digital constitutionalism" introduces the idea that intermediaries need to meet good governance standards to legitimately govern speech.[677] Various actors declared that self-regulation alone has proven insufficient to hold powerful commercial actors accountable.[678] Currently, content governance builds on the contractual terms of services that constitute a "take-it-or-leave-it" decision for individuals wanting to use intermediaries' services. Procedural limits to the possibility of defining and enforcing one's own policies could protect rights and due process, limit arbitrariness and, ultimately, strengthen rule of law.[679]

The current self-regulatory system also contributed to the massive information asymmetry, which precludes user agency. Civil society, academia and international organizations alike have suggested measures to empower individuals. These could include providing individualized interfaces to ensure individuals have control over what they see and options to choose between different approaches to content governance.[680]

---

[672] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 11.
[673] Au et al, Profiting from the Pandemic, p 8.
[674] UN Special Rapporteur on freedom of expression, A/HRC/38/35, para 46. Google/YouTube provides examples for its COVID-19 disinformation policies, see YouTube, COVID-19 policies.
[675] UN Special Rapporteur on freedom of expression, A/HRC/38/35, para 48.
[676] Ofcom, Use of AI in Online Content Moderation, p 43.
[677] Suzor, Lawless: the Secret Rules That Govern Our Digital Lives.
[678] For example, RFoM, SAIFE Policy Manual.
[679] Suzor, Lawless: the Secret Rules That Govern Our Digital Lives.
[680] EFF, Content Moderation is Broken, Let Us Count the Ways, p 5. Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 11 and Ofcom, Use of AI in Online Content Moderation, p 72.

This could include access to more diversity,[681] in order to proactively receive information beyond assumed or known interests, or curation based on moods or interests at a given moment.[682] Content governance policies could even be provided by third parties.[683] At the same time, however, such design options need to keep in mind that people may deliberately choose filter bubbles, especially if interfaces are not well designed or implemented.[684] While calls for customization are often framed around cultural norms regarding nudity, for example, it could be equally relevant in the context of disinformation. Interface options could not only provide individuals with options to adjust their exposure to diverging information, but also whether questionable information should rather be made less visible, be labeled, or annexed with related articles.

5.3.7 Human rights impact assessments

Due diligence procedures such as human rights impact assessments (HRIAs) can be essential tools for ensuring respect for human rights in content governance, including when tackling COVID-19 disinformation. HRIAs of policies and their implementation as well as the use of automated tools can help identify risks and address them before harm occurs (on the individual and collective level). Moreover, they can constitute the basis for both transparency and oversight.

Many jurisdictions oblige private actors to conduct environmental impact assessments to evaluate the (potential) effects of proposed projects, and some laws require data protection impact assessments. The EU GDPR, for example, requires an assessment prior to adopting an application with a "high risk" to an individual's rights, for instance "due to a systematic and extensive evaluation of personal aspects [...] based on automated processing".[685] Such assessments typically require evaluating necessity and proportionality. If they include publication obligations, they give individuals more control and can enhance accountability. The proposed EU AI Act also introduces mandatory impact assessments, just as the proposed U.S. Algorithmic Accountability Act.[686]

Some internet intermediaries announced or already undertook impact assessments, albeit with limited in scope.[687] So far, the findings were often not made publicly available or independently auditable, which weakens HRIAs' potential as a human rights safeguard.

---

[681] Judith Möller, Damian Trilling, Natali Helberger, Brams van Es, Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity, Information, Communication & Society, Volume 21, Item 4, pp 1-19, March 2018; http://dx.doi.org/10.1080/1369118X.2018.1444076.

[682] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 23.

[683] Gillespie, Custodians of the Internet, p 199.

[684] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 23.

[685] Art 35 para 3(a).

[686] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 145.

[687] Facebook, for example, did an assessment of the human rights impact of Facebook in Myanmar in 2018 after enormous public pressure, see Facebook, An Independent Assessment of the Human Rights Impact of Facebook in Myanmar, November 2018; https://about.fb.com/news/2018/11/myanmar-hria and in Indonesia, Sri Lanka, Cambodia, see Facebook, An Update on Facebook's Human Rights Work in Asia and Around the World, May 2020; https://about.fb.com/news/2020/05/human-rights-work-in-asia.

It often remained unclear to what extent identified risks resulted in mitigation efforts and ultimately avoided harm.

Given machine learning's adaptability, assessments of automated content governance may only be effective if they are undertaken periodically and not only ex ante. Data feeding back into the system may cause biases, distortions, or errors previously not identifiable. Flawed training data for machine-learning tools deployed in the context of disinformation, for example, may not be identified in a one-time assessment.[688] From a human rights perspective, HRIAs should thus include regular and systemic evaluations of policies and their (automated) enforcement regarding, inter alia, the accuracy, fairness, bias and discrimination, and their impact on privacy and security.

In order to ensure a broad and systemic evaluation of automated tools, the Council of Europe calls for democracy, rule of law and human rights impact assessments.[689] In the same vein, HRIAs should be holistic and include the conception, design, testing and deployment phases of automated tools.[690] Moreover, to be effective, consequences have to follow from HRIA findings. This would include that if mitigation measures are not sufficient to remedy identified human rights risks, the policy and/or automated tool should not be deployed.[691] For these reasons, guidance for genuine HRIAs call for accountability, enforceability and potential grievance mechanisms.[692]

Moreover, it is relevant who conducts HRIAs. To enhance transparency and public scrutiny, it would be particularly important that outcomes based on self-assessed HRIAs are disclosed.[693] Also, HRIAs could entail public comment processes, as is often the case in environmental law, to achieve a greater acceptance, to define the scope, and to ensure multi-stakeholder involvement in the decision process. This would result in the inclusion of valuable knowledge of affected groups, foster information exchange, and ensure that intermediaries are not advancing the interests of specific interest groups over those of the larger public.[694] Moreover, such inclusive processes could facilitate locally adapted implementations of impact assessments.[695]

---

[688] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 165f and RFoM, SAIFE Policy Manual.

[689] For guidance, see Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Policy Development Group, Human Rights, Democracy and Rule of Law Impact Assessment of AI systems, May 2021; https://rm.coe.int/cahai-pdg-2021-02-subworkinggroup1-ai-impact-assessment-v1-2769-4229-7/1680a1bd2d.

[690] Lorna McGregor, Daragh Murray, Vivian Ng, international Human Rights Law as a Framework for Algorithmic Accountability, Cambridge University Press, International and Comparative Law Quarterly, April 2019; http://dx.doi.org/10.1017/S0020589319000046.

[691] The UN HCHR calls for moratoriums for high risk AI systems, see UN High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/48/31, para 59(c). Also the EU AI Act proposal calls for a ban of certain AI applications.

[692] Guidance provided, e.g. by the Danish Institute for Human Rights, Guidance on Human Rights Impact Assessment of Digital Activities, November 2020; https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf and RFoM, SAIFE Policy Manual, p 47f.

[693] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 184 and p 187ff.

[694] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 185ff.

[695] Rabat Plan of Action, A/HRC/22/17/Add.4. See also UN Special Rapporteur on freedom of expression, hate speech, A/74/486, para 58(d) and (e).

5.3.8 Transparency

Various actors continually criticize the overall limited transparency in the digital ecosystem, including in the context of content governance and responses to online falsities and deception. Transparency, however, is a requirement for developing evidence-based responses. Information is needed to understand dissemination patterns of COVID-19 disinformation, as well as its dynamics, scale, and scope. Transparency would also be a precondition for assessing fact-checking and other responsive strategies to disinformation[696] and to the pandemic more broadly.[697] Access to intermediaries' data would allow an assessment of whether, and to what extent, disinformation establishes fragmentation online, and amplifies extremism and violence.[698] Open-source investigations would also enable evaluations on whether intermediaries are enhancing or algorithmically reinforcing deceptive behavior.[699]

A lack of access to data and transparency, to the contrary, prevents objective scrutiny of intermediaries' responses to disinformation, of their human rights impact, and of the effectiveness of mitigation measures. Opacity significantly disempowers users, as underlined by the UN Special Rapporteur on freedom of expression,[700] particularly as intermediaries regularly experiment with specific automated functionalities on their services, while users, the public, and regulators are left in the dark.[701] Furthermore, keeping automated content governance as "black boxes" may contribute to an assumption of neutrality or objective representation of in fact heavily customized, curated content.[702]

More recently, following public and regulatory pressure, most intermediaries provide transparency reports. Yet, these typically biannual reports only provide limited information such as the number of content takedowns.[703] Given the lack of harmonized standards of what is or should be included or how information should be provided, findings are regularly incomparable across the sector. Recent reports have also not contextualized the information provided with the COVID-19 pandemic.[704] Even specific COVID-19 reports, such as those based on the EU Code of Practice on Disinformation, have been assessed as insufficient and criticized for being outside the rule of law

---

[696] Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 16.

[697] For this reason, Facebook announced to share data for health research, see Facebook, Data for Good: New Tools to Help Health Researchers Track and Combat COVID-19; https://about.fb.com/news/2020/04/data-for-good. The Data for Good programme intends to academics and health agencies to study the impact of mobility patterns on infection rates.

[698] Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation, p 13.

[699] François, Actors, Behaviors, Content: A Disinformation ABC, p 5.

[700] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 80.

[701] Gorwa et al., Algorithmic content moderation, p 11.

[702] Haas, Freedom of the Media and Artificial Intelligence, p 2.

[703] Ranking Digital Ranking, 2020 Accountability Index. Ethan Zuckerman, I read Facebook's Widely Viewed Content Report. It's really strange. August 2021; https://ethanzuckerman.com/2021/08/18/facebooks-new-transparency-report-is-really-strange.

[704] Broadband Commission, Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression, p 193. In its only decision so far concerning a case related to COVID-19, the Facebook Oversight Board recommends that Facebook improves its transparency reporting on health misinformation content moderation, see FOB, COVID.

system.[705] Typically, transparency reports lack information about when automated tools are used, how they work, or the key criteria that underpin particular decisions and consequences, let alone provide aggregated data or accuracy rates.[706] They thus often fail to provide explainability, or auditability.[707] Although there is no one-size-fits-all approach to explainability,[708] the lack of meaningful transparency regarding automated tools impede replicability and thus oversight.[709]

Typically, transparency reports also provide no specific data on disinformation, such as how much users engage with content that is later debunked, or how automated tools are used to address it, and how successfully. Reports normally also lack information on public-private cooperation to detect and act upon falsities online.[710]

While certain intermediaries provide their users with some information beyond transparency reports, for example by enabling inquiries into why certain ads are shown, these possibilities and responses thereto remain limited and often do not include data on user classifiers, economic value, or the use of automation.[711]

Explainability and transparency to individuals, however, are essential to enable individual redress for automated decisions and a general public debate about policies.[712] For this reason, the GDPR and Council of Europe Convention 108+ include a right to explanation for solely automated processes, granting individuals the right to demand "meaningful information about the logic involved" in automated processing of data and thus "explanations" about automated content recommender systems.[713]

As long as such explainability is not provided, only intermediaries themselves can see and assess the full picture of false COVID-19 content or their governance decisions.[714] . The prevailing lack of transparency underpinning today's digital ecosystem prevents an independent assessment of the impact of automated curation on disinformation, on which criteria recommendations are derived, or whether, and to what extent, micro-targeting facilitates disinformation. This absence of quantitative and qualitative information may even enable the selling of data to the highest bidder for micro-targeting messages. It may also facilitate dark patterns nudging particular behavior.[715] Comprehensive transparency can thus be a first step to remedy for the information disorder. It would be particularly effective if incorporated into a wider regulatory structure to guarantee that intermediaries operate in the public interest.[716]

---

[705] Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, p 17.
[706] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348 para 84. Bloch, Automation in Moderation, p 88.
[704] McGregor et al., international Human Rights Law as a Framework for Algorithmic Accountability.
[708] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 33f.
[709] European Parliament, Study on regulating disinformation with artificial intelligence, p 18.
[710] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 81f.
[711] Facebook, for example, provides the possibility to ask why a particular ad is shown, see Facebook, How does Facebook decide which ads to show me?; https://www.facebook.com/help/794535777607370.
[712] For more information on accountability and redress, see chapter 5.3.9.
[713] See Article 13 para 2(f) and Recital 60 GDPR.
[714] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 183.
[715] UNESCO, Letting the Sun Shine In: Transparency and Accountability in the Digital Age, World Trends in Freedom of Expression and Media Development, May 2021; https://unesdoc.unesco.org/ark:/48223/pf0000377231, p 6.
[716] MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, p 5.

In general, a lack of transparency can be instrumentalized for disinformation, as evidenced by the Cambridge Analytica scandal.[717] While independent media and researchers investigated and exposed this scandal, a lack of transparency also significantly hampers independent scrutiny and investigations.[718]

Recent regulatory responses or proposals, such as the DSA or U.S. Social Media DATA Act, recognize the need for access to data for the research community.[719] Envisaged data-sharing frameworks could provide a valid basis for policies and break with intermediaries' current practice of rejecting disinformation research, including by ignoring internal findings,[720] legally threating advocacy groups,[721] and suspending accounts of researchers.[722]

Currently, intermediaries often refuse to share access to their data for independent assessments. They regularly base their refusals on arguments of complexity[723] or commercial intellectual property rights.[724] Frequently, privacy is invoked as a barrier to disclosing information. While the protection of privacy is an important aim, it should not be leveraged to avoid scrutiny and transparency, or creating barriers for competitors and new entrants.[725] Moreover, if intermediaries opaquely cooperate with governments to remove harmful content, it may not be credible if they at the same time invoke data protection concerns against efforts to increase public scrutiny.[726]

Besides, valid privacy concerns could be addressed, for example, through solid legal frameworks and a tiered approach with different transparency requirements for the general public, affected users, researchers, and regulators.[727] This would also avoid infringements of trade secrets or proprietary interests.[728]

While access to data is a precondition to better understand online disinformation and its broader social media impact, certain risks persist.[729] Far-reaching transparency, for example, can enable reverse engineering and thus undermine the effectiveness of

[714] The Guardian, The Cambridge Analytica Files, March 2018; https://www.theguardian.com/news/series/cambridge-analytica-files.
[715] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 80.
[719] Protocol, Issie Lapowsky and Ben Brody, Lawmakers want to force Big Tech to give researchers more data, May 2021; https://www.protocol.com/policy/social-media-data-act.
[720] The Facebook Files unveil the extent of internal research on harm that has been repeatedly ignored. See The Wall Street Journal, The Facebook Files. Internal oversight and accountability would also improve organizational behavior.
[721] POLITICO, Mark Scott, Campaigners wanted more transparency. Facebook threatened to sue, August 2020; https://www.politico.eu/article/facebook-transparency-privacy-matthias-spielkamp-algorithm-watch/?utm_source=POLITICO.EU&utm_campaign=281c8a7bd3-EMAIL_CAMPAIGN_2021_08_19_10_15&utm_medium=email&utm_term=0_10959edeb5-281c8a7bd3-190567583.
[722] Facebook took down accounts of NYU researchers (Ad Observatory) investigating advertising disinformation and analyzing which political ads arrive in which newsfeeds. See POLITICO, Mark Scott, Facebook's attempt to ban academics runs into trouble, August 2021; https://www.politico.eu/article/facebook-nyu-laura-edelson-political-ads.
[723] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 15.
[724] CDT, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, p 33f.
[725] Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 31.
[726] *Ibid*, p 31.
[727] MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, pp 17ff.
[728] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 187ff.
[729] Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy.

disinformation responses.[730] Previous data breaches also highlight the privacy concerns involved in any data sharing.[731]Moreover, the current infodemic has generally resulted in an increased monitoring of online behavior and content. Government access to personal information made publicly available would infringe individuals' right to privacy.[732] As anonymity is essential for freedom of opinion and expression, disclosing information about an individual in the absence of judicial warrants can thus problematic.[733] In this vein, the ECtHR identifies a chilling effect if personal data revealing certain opinions is not protected.[734] It is thus important to balance data access with strong privacy, data and due process protection, and safeguards against government surveillance.[735]

5.3.9 Redress, accountability, and independent oversight

Human rights due diligence and genuine good governance necessitates access to redress. The UN Special Rapporteur on freedom of expression clearly stated that individuals using services in the digital ecosystem are in fact rightholders, and not merely users dependent on intermediaries' rules shaping their speech and access to information.[736] In the context of automated moderation of COVID-19 disinformation, this means that intermediaries have to provide their users with adequate, easy-to-find, cost-free remedies for wrongful takedowns or deprioritization. While most intermediaries provide appeal mechanisms for removals, they typically do not provide redress for other interventions such as disinformation labels, demotions, or demonetization, or for other actions against inauthentic behavior.[737] Moreover, just as fact-checking is limited in some parts of the world,[738] so too are some appeal mechanisms only available in certain majority languages.[739] Ideally, however, redress would be provided for any interventions, arguably even for interface and design choices affecting individuals' rights.

In order to be effective, remedy mechanisms should ensure human review by someone with cultural and linguistic expertise, although this inevitably prolongs any grievance procedure. Automated moderation takes place at speed and scale, while redress is typically not scaleable.[740] Thus, as appeals are typically burdensome and

---

[730] Ofcom, Use of AI in Online Content Moderation, p 40.
[731] See, UN High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/48/31, para 14. Also, for example Cambridge Analytica.
[732] UN High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/39/29, para 6.
[733] In context of Germany's Enforcement Act, Van Hoboken et al., Regulating Disinformation in Europe: Implications for Speech and Privacy, p 29f.
[734] European Court of Human Rights, Catt v. United Kingdom, application no. 43514/15, judgment, 24 January 2019; http://hudoc.echr.coe.int/eng?i=001-189424, para 80.
[735] MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, p 9.
[736] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 84.
[737] *Ibid*, para 72.
[738] Meedan, 2020 Misinfodemic Report: COVID-19 in Emerging Economies; https://meedan.com/reports/2020-misinfodemic-report-covid-19-in-emerging-economies.
[739] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 72. For an analysis, see Online Censorship, How to appeal; https://onlinecensorship.org/resources/how-to-appeal.
[740] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 10.

lengthy, automated content governance of COVID-19 disinformation might lead to serious procedural-remedial concerns.[741]

For redress to be accessible, affected individuals need to be notified, as stipulated by the ECtHR,[742] including of how decisions impact them and on which criteria such decisions are based. Notifications should entail the reason for the individual decision, the margin of discretion, whether automation was involved, and how the specific context was analyzed.[743]

In order to tackle problematic speech on their services, intermediaries typically provide notice-and-takedown procedures. Yet, notification or reporting tools are regularly unavailable for false content. The question of who can report what kind of potentially problematic speech is thus crucially important for accountability mechanisms. Given the complexity of COVID-19 disinformation and potential misuse by users, not all intermediaries allow for in-product user report, but rather rely on trusted sources, flaggers or fact-checking organizations.[744]

A different, human rights-friendly way proposed to provide individuals with a possibility to report disinformation is a notice-and-notice procedure where a notification is forwarded to the content provider to enable a means of settling of the dispute.[745] Alternatively, it has been suggested that intermediaries consider notice-and-review mechanisms, where notified content is flagged for fact-checkers, providing for a regime against bad faith reporting. Consequently, notification and user reporting mechanisms could also be enabled for private messaging apps, while protecting end-to-end encryption.[746]

As much of intermediaries' content governance is not transparently disclosed, however, individuals regularly do not know when their content is subject to measures besides takedowns, and how their content is shared or accessed by others. For this reason, the Santa Clara Principles on Transparency and Accountability, for example, set out specific criteria for comprehensive disclosure and effective redress.[747]

---

[741] Bloch, Automation in Moderation, p 90f.

[742] European Court of Justice, Roman Zakharov v Russia, application no. 47143/06, judgment 4 December 2015; http://hudoc.echr.coe.int/eng?i=001-159324, para 161. With regard to Facebook, the German Federal Court in July 2021 ruled that notification of users is a necessary precondition for adequate remedies, see German Federal Court, judgment, Az. III ZR 179/20, 29 July 2021; http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=121741&pos=0&anz=1.

[743] Council of Europe, Algorithms and Human Rights, p 25.

[744] See, for example, Twitter, COVID-19 policies. TikTok, on the contrary, allows users to report falsities on their service, see TikTok, COVID-19 policies.

[745] European Parliament, Disinformation and propaganda, p 85.

[746] Forum on Information & Democracy, Working Group on Infodemics, Policy Framework, November 2020; https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf, p 92ff. For specific proposals on how to ensure encryption for such notifications, see p 102f. For more information on end-to-end encryptions and ways to moderate content on such platforms, see Dhanaraj Thakur, Guest Post, Mallory Knodel, Emma Llansó, Greg Nojeim and Caitlin Vogus, Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems, Center for Democracy & Technology, August 2021; https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems.

[747] Santa Clara Principles on Transparency and Accountability in Content Moderation, February 2018; https://santaclaraprinciples.org.

For instance, it would useful to develop clear standards on how affected users are notified and to provide them with reply and counter-notification opportunities in case their content is flagged as misleading. Remediation could involve reinstatements of content or public statements, or other forms of guarantees of non-repetition or compensation, depending on the harm caused. Moreover, an equivalent to class-action lawsuits could be considered given the collective human rights impact of automated content governance on the societal level.

Another challenge in the context of automated content governance decisions is to establish individual accountability. This could be linked to the person who originally was in a position to identify and prevent harm or mitigate a certain risk, for example in the context of HRIAs. The original developer, however, may not know about the automated tools' implementation and the person implementing an automated tool, on the other hand, may not be aware of the technical details. Other accountability systems such as product liability or corporate liability have thus been identified as more feasible.[748] If redress statistics, including reasons for re-installments or other corrective measures are disclosed for independent review, transparency could further strengthen accountability.[749]

As stated by the European Data Protection Board, online manipulation may be seen as a symptom of a general lack of accountability in the online environment.[750] Intermediaries design quasi-legal systems, their content governance thus needs to be accountable to democratic processes in order to be in be in line with human rights standards.[751] To ensure freedom of expression by design, accountability would need to be incorporated into the design, development and deployment of all technologies and practices.[752] Accountability could also be increased by introducing a Freedom of Expression Officer, comparable to the Compliance Officer envisaged by the EU DSA, combined with public and judicial oversight,[753] or public representatives overseeing content governance and intermediaries' general adherence to human rights from the inside, similar to Federal Reserve examiners for large banks.[754]

For broad accountability, strong remedies and genuine mitigation of existing individual and societal harms, independent oversight is essential. To date, despite some unilateral initiatives such as the Facebook Oversight Board with its semi-external complaints mechanism, there is no industry-wide or truly independent oversight.[755] Moreover, existing initiatives typically focus on individual content decisions with rather "precedential" aims, disregarding the broader context of curation and business models.

---

[748] Council of Europe, Algorithms and Human Rights, p 39. This approach is also rather used in the EU AI Act.
[749] Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, p 13ff.
[750] European Data Protection Board, Opinion 3/2018, Opinion on online manipulation and personal data, p 22.
[751] Bloch, Automation in Moderation, p 93.
[752] Llansó et al., Artificial Intelligence, Content Moderation, and Freedom of Expression, p 13.
[753] European Parliament, Disinformation and propaganda: 2021 update, p 121 and p 125.
[754] This proposal was made by Facebook whistleblower Frances Haugen. See The Wall Street Journal, The Facebook Files, Part 6: The Whistleblower, October 2021; https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-6-the-whistleblower/b311b3d8-b50a-425f-9eb7-12a9c4278acd?mod=series_facebookfiles.
[755] UN Special Rapporteur on freedom of expression, Disinformation, A/HRC/47/25, para 72.

Narrowing oversight mechanisms to questions of content removals alone frames discussions around disinformation and human rights implications of intermediaries' services in a narrow way that is convenient for them.

Self-regulation and self-assessment can provide important information and be practicable due to rapid technological changes. Yet, they can be seriously biased or misleading, and block true assessments of the impact of fragmented "publics" of information on individual and collective rights.[756] Given the substantial impact that responses to disinformation have on society, human rights-friendly content governance needs to be subject to objective external oversight.[757]

Oversight is particularly important in the context of automation, as the use of automated tools additionally risks interfering with individual rights, bypassing the rule of law and weakening democratic values.[758] Just as the GDPR requires additional scrutiny for high-risk activities, the UN Special Rapporteur on freedom of expression called on policy to require further oversight for automated tools such as independent audits, audit trails or requesting proofs of conformation to certain properties.[759]

Inspiration for independent audits, potentially funded separately from the industry or building on co-regulatory approaches, can be drawn from the Domain area, for example, where industry-shaped rules are based on a regulatory design.[760] Dual mechanisms with robust self-assessment followed by external, independent public review could be a useful way forward. Concerns could also be addressed through industry-wide social media councils as suggested by civil society and taken up by the UN Special Rapporteur on freedom of expression. As independent, participatory accountability mechanisms, they could consider appeals and provide general guidance in human rights-friendly content governance.[761]

6. Conclusion

While disinformation is no new phenomenon, the digital ecosystem presents a new terrain of scale and speed with vast amounts of user-generated content, new media and news-aggregation vectors, and informational gatekeeping powers of a few internet intermediaries. The uncertainties and serious health risks of the COVID-19 pandemic has further exacerbated disinformation's potential for harm. Recognizing the urgency to act, both states and intermediaries took unprecedented steps to address COVID-19 online disinformation.

---

[756] Nahmias et al., The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, p 190.
[757] *Ibid*, p 150.
[758] Elkin-Koren et al, Separation of Functions for AI, p 49.
[759] UN Special Rapporteur on freedom of expression, artificial intelligence, A/73/348, para 56f.
[760] European Parliament, Study on regulating disinformation with artificial intelligence, p 51; Nominet for the Domain area, For an analysis of existing practices of self- and co-regulation for content moderation, see Council of Europe, Guidance Note, Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory an co-regulatory mechanisms of content moderation, adopted by the Steering Committee for Media and Information Society (CDMSI), May 2021; https://rm.coe.int/content-moderation-en/1680a2cc18.
[761] ARTICLE 19, Social Media Councils; https://www.article19.org/social-media-councils, and ARTICLE 19, Side-stepping rights, p 37. See also UN Special Rapporteur on freedom of expression, A/HRC/38/35, para 63.

From a human rights perspective,[762] various state responses have been problematic,[763] and intermediaries largely responded in an inadequate manner.[764] Both sides' aim to find solutions at scale led to an ever-increasing deployment of automated tools, in spite of their context-blind and thus error-prone nature and their potential to infringe on freedom of expression. This paper thus concludes that an unwarranted use of automation to govern speech is regularly not in line with international human rights standards.

Today's digital ecosystem widely builds on manifold services and infrastructures by a few large internet intermediaries whose policies are informed by commercial interest. This sociotechnical context has potentially even amplified COVID-19 disinformation.[765] This paper therefore further concludes that – by inadequately responding to, and arguably even facilitating, disinformation – internet intermediaries often do not act in line with their commitments under the UNGP.

While focusing on automated responses to COVID-19 disinformation particularly, this paper highlights the close link of online disinformation to broader challenges of the digital ecosystem, and the risk that the latter pose to human rights and the rule of law. An online infrastructure that enables the targeting of individuals with precision, built on surveillance-based advertising, makes individuals vulnerable to manipulation and deception.

States have a positive obligation not only to protect but also to promote human rights on the individual as well as collective societal level. States need to enable an environment conducive to the freedom of opinion and freedom to seek, receive and impart all kinds of ideas and opinions. Accordingly, states have to address problematic speech such as COVID-19 disinformation that may otherwise have a chilling effect.

International human rights law provides for minimum safeguards in this regard. States should, as a basic principle, neither disseminate nor encourage false information, and refrain from outsourcing human rights protection to private actors such as intermediaries who have censorial power and bypass traditional checks and balances. Instead, they should take action, regulatory, if need be, to ensure intermediaries do not infringe on freedom of expression through a problematic or excessive use of automation to address COVID-19 disinformation that falls short of the three-part test of Article 19 of the ICCPR. Otherwise, states themselves fall short of their international obligations to protect human rights.

Consequently, states should ensure, on the substantive level, that intermediaries' content governance is in line with human rights standards. States should guarantee, on the procedural level, that transparency, accountability and oversight are implemented as part of human rights due diligence, and, on the remedial level, that adjective guarantees and redress mechanisms are provided, that users have agency and that individuals are

[762] This paper focuses on the international human rights framework and does not assess the application of existing national legislation regarding content moderation (such as the German Network Enforcement Act or Austrian Law on Communication Platforms) on COVID-19 disinformation or regulatory proposals (such as the EU Digital Services Act).
[763] For an assessment of state responses to COVID-19 disinformation, see chapter 5.2.
[764] For an assessment of internet intermediaries' responses to COVID-19 disinformation, see chapter 5.3.
[765] For online disinformation methods, see chapter 3.3.

empowered and their resilience strengthened. Regulatory responses should thus be process-oriented rather than focus on content.

In order to ensure human rights by default, states need to further address the sociotechnical context in addition to providing for a mitigation of human rights risks. Accordingly, to effectively tackle COVID-19 disinformation, it may be necessary to redefine advertising- and surveillance-based business models, address the concentration of power, and unbundle services provided by a few dominant internet intermediaries. Moreover, independent, quality journalism needs to be strengthened, as well as vibrant plurality, civic capacity, and bottom-up approaches to tackle, track and expose online falsehood. States should strengthen democratic scrutiny and empower local communities to engage with information flows more critically.

States' positive obligation to ensure a human rights-based approach naturally also concerns their own rulemaking, which should be evidence-based and involve multi-stakeholder participation. State interference with speech may only be justified when it is prescribed by law, necessary to achieve a legitimate aim (such as public health), proportionate to the aim pursued, and when an independent judicial oversight is guaranteed. Any policy action, by states or private actors, should be based on good governance. Addressing the challenges of the current unilateral and unaccountable content governance by intermediaries should not result in more state control over content or comparably unaccountable systems.

Overall, effective responses to COVID-19 online disinformation require various actors to join forces and undertake multifaceted efforts stretching across the entire information ecosystem. Acknowledging that there is no one-size-fits-all approach, responses to online disinformation should focus on identifying and debunking falsehood, and on addressing the main tactics of manipulation and the agents driving or profiting from disinformation, as well as those targeted by it or trusting in it. Responses need to be inclusive and tailored from a gender and intersectional perspective. Otherwise, they risk missing differences in how false content targets individuals differently, as well as how automated decisions affect different groups in society differently.[766]

Consequently, states and intermediaries should acknowledge that the online information landscape and its governance is complex and should resist simplistic narratives. Automation can provide a useful tool to analyze, detect and fact-check disinformation. The scale of online disinformation may even necessitate the use of some automated tools. Yet, automation is no silver bullet. Disinformation is always context-dependent, and falsity builds on meaning, which in itself is structured in ways beyond a single or simple fact or falsehood, often expressed through information but also emotions and signifiers of identity.[767] Automated responses inevitably miss or misidentify nuances, at scale, with a potentially greater impact on marginalized voices. It is thus essential that automated decision-making includes human involvement, human review,

[766] This is also linked to often restricted access to information for marginalized groups in society. Moreover, solely male authoritative facts and voices contradict the inclusivity of responses to both disinformation and the overall pandemic.

[767] UNESCO, Steering AI and Advanced ICTs for Knowledge Societies: A Rights, Openness, Access, and Multi-stakeholder Perspective, November 2019; https://unesdoc.unesco.org/ark:/48223/pf0000372132.locale=en, p 58.

and human reversibility, and provides appeal processes and independent audits. Otherwise, they fall short of human rights standards.

This paper illustrates that not only will COVID-19 have a lasting impact on the global digital community, but also the crisis-adjusted content governance of internet intermediaries. Their responses to COVID-19 disinformation may constitute a turning point as intermediaries take on ever increasing editorial roles that may, once tested in a crisis context, turn permanent. Increased editorial action or legally mandated media-like responsibilities will affect individuals' and societies' access to relevant and trusted information – and must go hand in hand with strong human rights safeguards.

7. List of references

7.1. Bibliography

Access Now, #KeepItO: Fighting internet shutdowns around the world; https://www.accessnow.org/keepiton

Access Now, Fighting Misinformation and Defending Free Expression During COVID-19: Recommendations for States, April 2020; https://www.accessnow.org/cms/assets/uploads/2020/04/Fighting-misinformation-and-defending-free-expression-during-COVID-19-recommendations-for-states-1.pdf

Access Now, Thailand: Stop Weaponizing 'COVID-19' to Censor Information "Causing Fear" and Crack Down on Media and Internet Service Provider, August 2021; https://www.accessnow.org/cms/assets/uploads/2021/08/Joint-Statement-Thailand-Regulation29-COVID19-August2021-FINAL.docx.pdf

Al Jazeera, Iran: Over 700 dead after drinking alcohol to cure coronavirus, April 2020; https://www.aljazeera.com/news/2020/4/27/iran-over-700-dead-after-drinking-alcohol-to-cure-coronavirus

Susie Alegre, Rethinking Freedom of Thought for the 21st Century, Doughty Street Chambers, April 2020, European Human Rights Law Review, 2017, Issue 3; https://susiealegre.com/wp-content/uploads/2020/04/Alegre%20from%202017_EHRLR_Issue_3_Print_final_0806%5B6745%5D.pdf

AlgorithmWatch, Automating Society Report 2020, Life in the automated society: How automated decision-making systems became mainstream, and what to do about it, September 2020; https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf

Julia Angwin and Hannes Grassegger, Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children, ProPublica, June 2017; https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, February 2018; https://ec.europa.eu/newsroom/article29/items/612053/en.

ARTICLE 19, Viral Lies: Misinformation and the Coronavirus, Policy Brief, March 2020; https://www.article19.org/wp-content/uploads/2020/03/Coronavirus-briefing.pdf

ARTICLE 19, Social Media Councils; https://www.article19.org/social-media-councils

ARTICLE 19, Side-stepping rights, June 2018; https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf

Evelyn Aswad, Losing the Freedom to Be Human, Columbia Human Rights Law Review, Vol. 53, February 2020; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635701

Yung Au, Philip N. Howard, Project Ainita, Profiting from the Pandemic: Moderating COVID-19 Lockdown Protest, Scam, and Health Disinformation Websites, COVID-19 Series, Oxford Internet Institute, November 2020; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2020/12/Profiting-from-the-Pandemic-v8-1.pdf

Avaaz, Left Behind: How Facebook is neglecting Europe's infodemic, April 2021; https://secure.avaaz.org/campaign/en/facebook_neglect_europe_infodemic

Avaaz, Facebook's Algorithm: A Major Threat to Public Health, August 2020 https://secure.avaaz.org/campaign/en/facebook_threat_health

Judit Bayer, Natalija Bitiukova, Petra Bárd, Judit Szakács, Alberto Alemanno, Erik Uszkiewicz, Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States, European Parliament, study requested by the LIBE committee, February 2019; https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf

Judit Bayer, Bernd Holznagel, Katarzyna Lubianiec, Adela Pintea, Josephine B. Schmitt, Judit Szakacs, Erik Uszkiewicz, Disinformation and propaganda: impact on the functioning of the rule of law in the EU and its Member States – 2021 update, European Parliament, study requested by the INGE committee, April 2021; https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653633/EXPO_STU(2021)653633_EN.pdf

BBC, Coronavirus: Outcry after Trump suggests injecting disinfectant as treatment, April 2020; https://www.bbc.com/news/world-us-canada-52407177

BBC, Raïssa Ioussouf, Madagascar president's herbal tonic fails to halt Covid-19 spike, August 2020; https://www.bbc.com/news/world-africa-53756752

Bellingcat, How Coronavirus Scammers Hide On Facebook And YouTube, March 2020; https://www.bellingcat.com/news/rest-of-world/2020/03/19/how-coronavirus-scammers-hide-on-facebook-and-youtube

Bellingcat, Investigating Coronavirus Fakes And Disinfo? Here Are Some Tools For You, March 2020; https://www.bellingcat.com/resources/2020/03/27/investigating-coronavirus-fakes-and-disinfo-here-are-some-tools-for-you

Wolfgang Benedek, Matthias C. Kettemann, Freedom of Expression and the Internet (2nd edition), September 2020

Walter Berka, Christina Binder, Benjamin Kneihs, Die Grundrechte, Grund- und Menschenrechte in Österreich, 2. Auflage, Verlag Österreich, November 2019

Berkman Klein Center for Internet & Society at Harvard University, Lumen Project, https://www.lumendatabase.org

Hannah Bloch-Wehba, Automation in Moderation, Cornell International Law Journal, Volume 53, Issue 1, pp 42-55, Texas A&M University School of Law, March 2020; https://scholarship.law.tamu.edu/facscholar/1448

Blackbird.AI, COVID-19 (Coronavirus) Disinformation Report, Volume 4.0, April 2021; https://www.blackbird.ai/reports

Kalina Bontcheva and Julie Posetti (ed.), Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression, Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet', International Telecommunication Union and UNESCO, September 2020; https://en.unesco.org/sites/default/files/8_challenges_and_recommended_actions_248_266_balancing_act_disinfo.pdf

Samantha Bradshaw, Hannah Bailey, Philip N. Howard, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, Computational Propaganda Research Project, Oxford Internet Institute, January 2021; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/02/CyberTroop-Report20-Draft9.pdf

Deborah Brown, Big Tech's Heavy Hand Around the Globe, Human Rights Watch, September 2020; https://www.hrw.org/news/2020/09/08/big-techs-heavy-hand-around-globe

Talha Burki, The online anti-vaccine movement in the age of COVID-19, The Lancet, Digital Health, Volume 2, Issue 10, October 2020; https://doi.org/10.1016/S2589-7500(20)30227-2

Paul Butcher, COVID-19 as a turning point in the fight against disinformation, Nature Electronics, Volume 4, pp 7-9, January 2021; https://doi.org/10.1038/s41928-020-00532-2

BuzzFeed News, Craig Silvermann, Ryan Mac, Pranav Dixit, "I Have Blood on My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation, September 2020; https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo

Dominiko Bychawska-Siniarska, Council of Europe, Protecting the Right to Freedom of Expression Under the European Convention on Human Rights, A handbook for legal practitioners, July 2017; https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814

Alex Campbell, How Data Privacy Laws Can Fight Fake News, Just Security, August 2019; https://www.justsecurity.org/65795/how-data-privacy-laws-can-fight-fake-news

Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches, Data & Society, November 2018, https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

Center for Countering Digital Hate (CCDH), The Disinformation Dozen, Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers; https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_b7cedc0553604720b7137f8663366ee5.pdf

Center for Democracy & Technology, COVID-19 Content Moderation Research, April 2020; https://cdt.org/insights/covid-19-content-moderation-research-letter

Center for Democracy & Technology, Feedback to the EU Commission proposal on Artificial Intelligence ("AI Act"), F2665242, August 2021; https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665242_en

Man-piu Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, Dolores Albarracin, Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation, Psychological Science, Vol. 28, Issue 11, pp. 1531-1546, May 2017; https://doi.org/10.1177/0956797617714579

Katherine Clayton, Spencer Blair, Jonathan A. Busam et al, Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media, Political Behavior, Vol. 42, Issue 4, pp 1073-1095, December 2020; https://doi.org/10.1007/s11109-019-09533-0

Coalition to Fight Digital Deception, Trained for Deception: How Artificial Intelligence Fuels Online Disinformation, September 2021; https://assets.mofoprod.net/network/documents/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation_T2pk9Wj.pdf

Nithin Coca, Disinformation from China floods Taiwan's most popular messaging app, Coda Story, October 2020; https://www.codastory.com/authoritarian-tech/taiwans-messaging-app

Carme Colomina, Héctor Sánchez Margalef, Richard Youngs, The impact of disinformation on democratic processes and human rights in the world, European Parliament, study requested by the DROI subcommittee, April 2021; https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf

CNN, Claire Duffy, For misinformation peddlers on social media, it's three strikes and you're out. Or five. Maybe more, September 2021, https://edition.cnn.com/2021/09/01/tech/social-media-misinformation-strike-policies/index.html

Committee on Economic, Social and Cultural Rights, Twenty-second Session, General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12), E/C.12/2005/4, https://www2.ohchr.org/english/bodies/cescr/docs/gc14

Committee to Protect Journalists (CPJ), Amid COVID-19, the prognosis for press freedom is dim: Here are 10 symptoms to track, June 2020; https://cpj.org/reports/2020/06/covid-19-here-are-10-press-freedom-symptoms-to-track

#CoronaVirusFacts Alliance, Poynter, https://www.poynter.org/coronavirusfactsalliance

Corporate Europe Observatory, The lobby network: Big Tech's web of influence in the EU, August 2021; https://corporateeurope.org/en/2021/08/lobby-network-big-techs-web-influence-eu

Council of Europe, Parliamentary Assembly, Resolution 1970 on "Internet and Politics: the impact of new information and communication technology on democracy", January 2014; https://pace.coe.int/en/files/20329

Council of Europe, Parliamentary Assembly, Resolution 2143 on "Online media and journalism: challenges and accountability", January 2017; https://pace.coe.int/en/files/23237

Council of Europe, Parliamentary Assembly, Resolution 2217, April 2018; https://assembly.coe.int/nw/xml/XRef/Xref-XML2HTML-en.asp?fileid=24762&lang=en

Council of Europe, Committee of experts on internet intermediaries (MSI-NET), Study on the Human Rights Dimension of Automated Data Processing Techniques (in particular Algorithms) and Possible Regulatory Implications, MSI-NET(2016)06 rev3 final, October 2017; https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a

Council of Europe, Algorithms and Human Rights: Study on the human rights dimension of automated data processing techniques and possible regulatory implications, Council of Europe study DGI(2017)12, prepared by the committee of experts on internet intermediaries (MSI-NET), March 2018; https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, 1337th meeting of the Ministers' Deputies, Decl(13/02/2019)1, February 2019; https://search.coe.int/cm/pages/result_details.aspx?ObjectId=090000168092dd4b

Council of Europe, Press freedom must not be undermined by measures to counter disinformation about COVID-19, April 2020; https://www.coe.int/en/web/commissioner/-/press-freedom-must-not-be-undermined-by-measures-to-counter-disinformation-about-covid-19

Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Feasibility Study, CAHAI(2020)23, December 2020; https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

Council of Europe, Guidance Note, Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory an co-regulatory mechanisms of content moderation, adopted by the Steering Committee for Media and Information Society (CDMSI), May 2021; https://rm.coe.int/content-moderation-en/1680a2cc18

Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Policy Development Group, Human Rights, Democracy and Rule of Law Impact Assessment of AI systems, May 2021; https://rm.coe.int/cahai-pdg-2021-02-subworkinggroup1-ai-impact-assessment-v1-2769-4229-7/1680a1bd2d

CSET, Ben Buchanan, Andrew Lohn, Micah Musser and Katerina Sedova, Truth, Lies, and Automation: How Language Models Could Change Disinformation, May 2021; https://cset.georgetown.edu/publication/truth-lies-and-automation

Danish Institute for Human Rights, Guidance on Human Rights Impact Assessment of Digital Activities, November 2020; https://www.humanrights.dk/sites/humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf

Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O. S. Vaz de Melo and Fabricio Benevenuto, Can WhatsApp Counter Misinformation by Limiting Message Fowarding?, September 2019; https://arxiv.org/pdf/1909.08740.pdf

Alexandre De Streel, Elise Defreyne, Hervé Jacquemin, Michèle Ledger, Alejandra Michel, Alessandra Innesti, Marion Goubet, Dawid Ustowski, Online Platforms' Moderation of Illegal Content Online: Laws, Practices and Options for Reform, European Parliament, study requested by the IMCO committee, June 2020; https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf

Joan Donovan, The Media Manipulation Casebook, Cloaked Science, 2020; https://mediamanipulation.org/definitions/cloaked-science

Domo, Data Never Sleeps 8.0, April 2020; https://www.domo.com/learn/infographic/data-never-sleeps-8

Christopher Dornan, Scientific Disinformation In a Time of Pandemic, Public Policy Forum, June 2020; https://ppforum.ca/publications/science-disinformation-in-a-time-of-pandemic

Evelyn Douek, The Oversight Board Moment You Should've Been Waiting For: Facebook Responds to the First Set of Decisions, February 2021; https://www.lawfareblog.com/oversight-board-moment-you-shouldve-been-waiting-facebook-responds-first-set-decisions

Natasha Duarte, Emma Llansó, Anna Loup, Mixed Messages? The Limits of Automated Social Media Content Analysis, Center for Democracy & Technology, November 2017; https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

EDRi, Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal, January 2021; https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal

Lilian Edwards and Michael Veale, Slave to the Algorithm? What a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For, Duke Law & Technology Review, Vol. 16, pp 18-84, December 2017; https://scholarship.law.duke.edu/dltr/vol16/iss1/2

Electronic Frontier Foundation (EFF), Cory Doctorow, Facebook's Secret War on Switching Costs, August 2021; https://www.eff.org/deeplinks/2021/08/facebooks-secret-war-switching-costs

Niva Elkin-Koren, Maayan Perel, Separation of Functions for AI: Restraining Speech Regulation by Online Platforms, Lewis & Clark Law Review, Vol. 24, Issue 3, pp. 857-898, August 2020; https://dx.doi.org/10.2139/ssrn.3439261

ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice, May 2020; https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf

EU Disinfo Lab, Misogyny and Misinformation: An Analysis of Gendered Disinformation Tactics During the DOCID-19 Pandemic, December 2002; https://www.disinfo.eu/publications/misogyny-and-misinformation%3A-an-analysis-of-gendered-disinformation-tactics-during-the-covid-19-pandemic

European Commission, Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling COVID-19 disinformation – Getting the facts right, JOIN(2020) 8 final, June 2020; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008

European Commission, Political advertising – improving transparency; https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12826-Transparency-of-political-advertising_en

European Commission, Shaping Europe's digital future: Code of Practice on Disinformation, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

European Commission, Monthly Reports on Fighting COVID-19 Disinformation; https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en (Results of working with

platforms) and https://digital-strategy.ec.europa.eu/en/news/coronavirus-disinformation-online-platforms-take-new-actions-and-call-more-players-join-code

European Commission, Communication from the Commission to the European Parliament and the Council, Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition – two years of application of the General Data Protection Regulation, June 2020; COM(2020) 264 final; https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0264&from=EN

European Court of Auditors, Special Report on Disinformation affecting the EU: tackled but not tamed, June 2021; https://www.eca.europa.eu/Lists/ECADocuments/SR21_09/SR_Disinformation_EN.pdf

European Court of Human Rights, Guide on Article 10 of the European Convention on Human Rights, Freedom of Expression, December 2020; https://www.echr.coe.int/documents/guide_art_10_eng.pdf

European Data Protection Board, Opinion 3/2018 on online manipulation and personal data, March 2018; https://edps.europa.eu/sites/default/files/publication/18-03-19_online_manipulation_en.pdf

European Data Protection Supervisor (EDPS), Opinion 1/2021 on the Proposal for a Digital Services Act, February 2021; https://edps.europa.eu/system/files/2021-02/21-02-10-opinion_on_digital_services_act_en.pdf

European Parliament, Opinion by the Committee on Civil Liberties, Justice and Home Affairs (LIBE), 2020/0361(COD), July 2021; https://www.europarl.europa.eu/doceo/document/LIBE-AD-692898_EN.pdf

EUROPOL, Catching the Virus: Cybercrime, Disinformation and the COVID-19 Pandemic, April 2020; https://www.europol.europa.eu/publications-documents/catching-virus-cybercrime-disinformation-and-covid-19-pandemic

EUvsDisinfo, https://euvsdisinfo.eu/category/blog/coronavirus

EUvsDisinfo, Kremlin Disinformation Impedes Russian Vaccination Efforts, March 2021; https://euvsdisinfo.eu/attacking-the-west-putting-russians-in-danger

EUvsDisinfo, Dizzy by Vaccine Disinfo - Capturing Vaccines Rollout in The EU's Neighbourhood and Russia, March 2021 https://euvsdisinfo.eu/dizzy-by-vaccine-disinfo-capturing-vaccines-rollout-in-the-eus-neighbourhood-and-russia

EUvsDisinfo, EEAS Special Report Update: Short Assessment of Narratives and Disinformation Around the COVID-19 Pandemic, April 2021; https://euvsdisinfo.eu/eeas-special-report-update-short-assessment-of-narratives-and-disinformation-around-the-covid-19-pandemic-update-december-2020-april-2021

Facebook, COVID-19 policies, https://about.fb.com/news/2020/03/combating-covid-19-misinformation; https://www.facebook.com/help/230764881494641 and https://about.fb.com/news/tag/covid-19

Facebook, Reaching Billions of People With COVID-19 Vaccine Information, February 2021; https://about.fb.com/news/2021/02/reaching-billions-of-people-with-covid-19-vaccine-information

Facebook, Data for Good: New Tools to Help Health Researchers Track and Combat COVID-19; https://about.fb.com/news/2020/04/data-for-good

Facebook, How does Facebook decide which ads to show me?; https://www.facebook.com/help/794535777607370

Facebook, Timeline: Taking action to combat misinformation, polarization, and dangerous organization, April 2021; https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Efforts-to-Combat-Misinformation-Polarization-and-Dangerous-Organizations.pdf

Facebook Oversight Board, Case decision 2020-006-FB-FBR, January 2021; https://oversightboard.com/news/325131635492891-oversight-board-overturns-facebook-decision-case-2020-006-fb-fbr

Facebook, Facebook's Response to the Oversight Board's First Set of Recommendations, February 2021; https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations

Facebook, How We're Taking Action Against Vaccine Misinformation Superspreaders, August 2021; https://about.fb.com/news/2021/08/taking-action-against-vaccine-misinformation-superspreaders

Facebook, Taking Action Against People Who Repeatedly Share Misinformation, May 2021; https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation

Facebook, Updating the Values That Inform Our Community Standards, September 2019; https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standard

Facebook, An Independent Assessment of the Human Rights Impact of Facebook in Myanmar, November 2018; https://about.fb.com/news/2018/11/myanmar-hria

Facebook, An Update on Facebook's Human Rights Work in Asia and Around the World, May 2020; https://about.fb.com/news/2020/05/human-rights-work-in-asia

D.J. Flynn, Brandan Nyhan, Jason Reifler, The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics, Political Psychology, Vol. 38, Issue S1, pp. 127-150, January 2017; https://doi.org/10.1111/pops.12394

Forum on Information & Democracy, Working Group on Infodemics, Policy Framework, November 2020; https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf

Camille François, Actors, Behaviors, Content: A Disinformation ABC, Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses, Transatlantic Working Group on Content Moderation Online and Freedom of Expression, September 2019; https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf

Camille François, Brookings Podcast on COVID-19 and the ABCs of disinformation, https://www.brookings.edu/techstream/podcast-camille-francois-on-covid-19-and-the-abcs-of-disinformation

Freedom House, Information Isolation: Censoring the COVID-19 Outbreak, March 2021; https://freedomhouse.org/report/report-sub-page/2020/information-isolation-censoring-covid-19-outbreak

Freedom House, Freedom on the net report 2021, September 2021; https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech#Internet

Freedom Online Coalition (FOC), Joint Statement on Spread of Disinformation Online, November 2020; https://freedomonlinecoalition.com/wp-content/uploads/2021/05/FOC-Joint-Statement-on-Spread-of-Disinformation-Online.pdf

Full Fact, Report on Facebook's Third-Party Fact-Checking programme, December 2020; https://fullfact.org/blog/2020/dec/full-fact-publishes-new-report-on-facebooks-third-party-fact-checking-programme

Daniel Funke, Daniela Flamini, A guide anti-misinformation actions around the world, Poynter, August 2019; https://www.poynter.org/ifcn/anti-misinformation-actions

Maximilian Gahnz, Katja T. J. Neumann, Philipp C. Otte, Bendix J. Sältz, Kathrin Steinbach, Breaking the News? Politische Öffentlichkeit und die Regulierung von Medienintermediären, Friedrich Ebert Stiftung, February 2021; http://library.fes.de/pdf-files/a-p-b/17429.pdf

Ryan J. Gallagher, Larissa Doroshenki, Sarah Shugars, David Lazer, Brooke Foucault Welles, Sustained Online Amplification of COVID-19 Elites in the United States, social media + society, June 2021; https://doi.org/10.1177/20563051211024957

German Federal Court, judgment, Az. III ZR 179/20, 29 July 2021; http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=121741&pos=0&anz=1

Amel Ghani, Sadaf Khan, Media Matters for Democracy, Disorder in the newsroom: the media's perceptions and response to the infodemic, December 2020; https://drive.google.com/file/d/1nOgwtFRH5hEtlRBYcayY5aAjEEm8aWBQ/view

Dipayan Ghosh and Ben Scott, Digital Deceit: The Technologies Behind Precision Propaganda on the Internet, New America, January 2018; https://www.newamerica.org/pit/policy-papers/digitaldeceit

Tarleton Gillespie, Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, January 2018

Github, Conversation AI, https://conversationai.github.io

Gizmodo, Tom McKay, Internal Facebook Documents Show How Badly It Fumbled the Fight Against Anti-Vaxxers: ReportSeptember 2021; https://gizmodo.com/internal-facebook-documents-show-how-badly-it-fumbled-t-1847696500

GDI, Research Brief: Ad tech fuels disinformation sites in Europe, March 2020; https://disinformationindex.org/wp-content/uploads/2020/03/GDI_Adtech_EU.pdf

GDI, Ad-funded COVID-19 Disinformation: Money, Brands and Tech, July 2020; https://disinformationindex.org/wp-content/uploads/2020/07/GDI_Ad-funded-COVID-19-Disinformation-1.pdf

GDI, Popular Brands Appearing Next to COVID-19 Anti-vaccination Disinformation, February 2021; https://disinformationindex.org/wp-content/uploads/2021/02/Feb_11_2021-DisinfoAds-_EU_COVID-19_AntiVaxx.pdf

Global Network Initiative (GNI), Implementation Guidelines; https://globalnetworkinitiative.org/implementation-guidelines

GNI, Principles on Freedom of Expression and Privacy; https://globalnetworkinitiative.org/gni-principles

Google/YouTube, COVID-19 policies, https://support.google.com/youtube/answer/9891785?hl=en, https://support.google.com/youtube/answer/9803260?hl=en and https://support.google.com/youtube/answer/9803260?hl=en#:~:text=If%20your%20content%20receives%20a,is%20now%20eligible%20for%20monetization

Google/YouTube, Protecting our extended workforce and the community, March 2020; https://blog.youtube/news-and-events/protecting-our-extended-workforce-and

Google/YouTube, Managing harmful vaccine content on YouTube, September 2021; https://blog.youtube/news-and-events/managing-harmful-vaccine-content-youtube/?utm_source=POLITICO.EU&utm_campaign=222ef59fd6-EMAIL_CAMPAIGN_2021_09_30_11_37&utm_medium=email&utm_term=0_10959edeb5-222ef59fd6-190567583

Robert Gorwa, Reuben Binns, Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big Data & Society, February 2020; https://doi.org/10.1177/2053951719897945

Andy Guess, Kevin Aslett, Joshua Tucker, Richard Bonneau, Jonathan Nagler, Cracking Open the News Feed,: Exploring What U.S. Facebook Users See and Share with Large-Scale Platform Data, Journal of Quantitative Description: Digital Media, Volume 1, April 2021; https://doi.org/10.51685/jqd.2021.006

Julia Haas, Freedom of the Media and Artificial Intelligence, Global Conference for Media Freedom, November 2020; https://www.international.gc.ca/campaign-campagne/assets/pdfs/media_freedom-liberte_presse-2020/policy_paper-documents_orientation-ai-ia-en.pdf

Natali Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power, Digital Journalism, Volume 6, Issue 6, pp 842-854, July 2020; https://doi.org/10.1080/21670811.2020.1773888

Philip Howard, Rasmus Kleis Nielson, Nic Newman, J. Scott Brannan, The COVID-19 'infodemic': what does the misinformation landscape look like and how can we respond?, Oxford Internet Institute, April 2020; https://www.oii.ox.ac.uk/blog/the-covid-19-infodemic-what-does-the-misinformation-landscape-look-like-and-how-can-we-respond

Human Rights Committee of the International Covenant on Civil and Political Rights, 102nd session, General comment No. 34, CCPR/C/GC/34, https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

Human Rights Council, Seventeenth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, 16 May 2011, A/HRC/17/27; https://undocs.org/A/HRC/17/27

Human Rights Council, Twentieth session, Resolution adopted by the Human Rights Council on 16 July 2012, The promotion, protection and enjoyment of human rights on the Internet, July 2012, A/HRC/RES/20/8; https://undocs.org/A/HRC/RES/20/8

Human Rights Council, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, A/HRC/22/17/Add.4; Appendix, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, January 2013; https://undocs.org/A/HRC/22/17/Add.4

Human Rights Council, Twenty-ninth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 22 May 2015, A/HRC/29/32; https://www.undocs.org/A/HRC/29/32

Human Rights Council, Thirty-second session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 11 May 2016, A/HRC/32/38; https://undocs.org/A/HRC/32/38

Human Rights Council, Thirty-fifth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, 30 March 2017, A/HRC/35/22; https://www.undocs.org/A/HRC/35/22

Human Rights Council, Thirty-eighth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye, A/HRC/38/35, April 2018; https://undocs.org/A/HRC/38/35

Human Rights Council, Thirty-ninth session, Report of the United Nations High Commissioner for Human Rights, The right to privacy in the digital age, A/HRC/39/29, August 2018; https://undocs.org/A/HRC/39/29

Human Rights Council, Thirty-ninth session, Report of the independent international fact-finding mission on Myanmar, 12 September; https://undocs.org/en/A/HRC/39/64 and Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar, A/HRC/39/CRP.2, September 2018; https://undocs.org/A/HRC/39/CRP.2

Human Rights Council, Forty-first session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "surveillance and human rights", David Kaye, A/HRC/41/35, May 2019; https://undocs.org/A/HRC/41/35

Human Rights Council, Forty-first session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "Disease pandemics and the freedom of opinion and expression", David Kaye, A/HRC/44/49, April 2020; https://undocs.org/A/HRC/44/49

Human Rights Council, Forty-fourth session, Resolution adopted by the Human Rights Council on 16 July 2020, Freedom of opinion and expression, A/HRC/44/12, July 2020; https://undocs.org/en/A/HRC/RES/44/12

Human Rights Council, Forty-six session, Minority issues, Report of the Special Rapporteur on minority issues, Fernand de Varennes, A/HRC/46/57, March 2021; https://undocs.org/A/HRC/46/57

Human Rights Council, Forty-seventh session, Disinformation and freedom of opinion and expression, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, A/HRC/47/25, April 2021; https://undocs.org/A/HRC/47/25

Human Rights Council, Forty-eight session, Report of the United Nations High Commissioner for Human Rights on "The right to privacy in the digital age", Michelle Bachelet, A/HRC/48/31, September 2021; https://www.ohchr.org/EN/Issues/DigitalAge/Pages/cfi-digital-age.aspx

Human Rights Watch, "Video Unavailable": Social Media Platforms Remove Evidence of War Crimes, September 2020; https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes

Independent High level Expert Group on Fake News and Online Disinformation, A multi-dimensional approach to disinformation, Report for the European Commission, March 2018; https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation

Instagram, Helping People Stay Safe and Informed about COVID-19 Vaccines, March 2021; https://about.instagram.com/blog/announcements/continuing-to-keep-people-safe-and-informed-about-covid-19

Institute of Electrical and Electronics Engineers (IEEE), 7000™-2021 Standards: Addressing Ethical Concerns During Systems Design, September 2021; https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html?utm_source=POLITICO.EU&utm_campaign=52a2ab7f3a-EMAIL_CAMPAIGN_2021_09_15_08_59&utm_medium=email&utm_term=0_10959edeb5-52a2ab7f3a-190567583

Institute for Strategic Dialogue (ISD), Covid-19 Disinformation Briefing No. 1, March 2020; https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-1

Institute for Strategic Dialogue (ISD), Covid-19 Disinformation Briefing No. 2, April 2020; https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-2

Institute for Strategic Dialogue (ISD), Disinformation Overdose: A study of the Crisis of Trust among Vaccine Sceptics and Anti-Vaxxers, July 2021; https://www.isdglobal.org/isd-publications/disinformation-overdose-a-study-of-the-crisis-of-trust-among-vaccine-sceptics-and-anti-vaxxers

International Press Institute, Rush to pass 'fake news' laws during Covid-19 intensifying global media freedom challenges, October 2020; https://ipi.media/rush-to-pass-fake-news-laws-during-covid-19-intensifying-global-media-freedom-challenges

International Center for Journalists (ICFJ) and Tow Center for Digital Journalism at Columbia University, Journalism and the Pandemic, October 2020; https://www.icfj.org/our-work/journalism-and-pandemic-survey

ICFJ and UNESCO, Global Study: Online Violence Against Women Journalists, November 2020; https://www.icfj.org/our-work/icfj-unesco-global-study-online-violence-against-women-journalists

ICFJ, Maria Ressa: Fighting on Onslaught of Online Violence – A big data analysis, March 2021; https://www.icfj.org/our-work/maria-ressa-big-data-analysis

Cherilyn Ireton and Julie Posetti, UNESCO, Journalism, 'Fake News' & Disinformation, Handbook for Journalism Education and Training, UNESCO Series on Journalism Education, November 2018, https://unesdoc.unesco.org/ark:/48223/pf0000265552_eng

Nina Jankowicz, How Disinformation Became a New Threat to Women, Female politicians and other high profile women face a growing threat from sexualized disinformation, Coda

Story, December 2017; https://www.codastory.com/disinformation/how-disinformation-became-a-new-threat-to-women

Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, March 2017; https://www.osce.org/files/f/documents/6/8/302796.pdf

Joint industry statement on COVID-19 from Microsoft, Facebook, Google, LinkedIn, Reddit, Twitter and YouTube, March 2020; https://twitter.com/Microsoft/status/1239703041109942272

Joint Statement COVID-19: Governments must promote and protect access to and free flow of information during pandemic, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, March 2020; https://www.osce.org/representative-on-freedom-of-media/448849

Joint Declaration on Freedom of Expression and Elections in the Digital Age, UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, April 2020; https://www.osce.org/files/f/documents/9/8/451150_0.pdf

Vladan Joler, New Extractivism, 2020, https://extractivism.online

Kate Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, International Law Programme, Chatham House, November 2019; https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf

David Kaye, The Clash Over Regulating Online Speech, SLATE, June 2019; https://slate.com/technology/2019/06/social-media-companies-online-speech-america-europe-world.html

Daphne Keller, Internet Platforms, Observations on Speech, Danger, and Money, National Security, Technology, and Law, Aegis Series Paper No. 1807, Hoover Institution, June 2018;
https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf

Kate Klonick, The New Governors of Speech: The People, Rules, and Processes Governing Online Speech, Harvard Law Review, Vol. 131, No. 6, pp. 1598-1620, April 2018; https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech

Aleksi Knuutila, Aliaksandr Herasimenka, Hubert Au, Jonathan Bright, Philip N. Howard, COVID-Related Misinformation on YouTube: The Spread of Misinformation Videos on Social Media and the Effectiveness of Platform Policies, Oxford Internet Institute, September 2020; https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/09/Knuutila-YouTube-misinfo-memo-v1.pdf

Mihalis Kritikos, Tackling Mis- and Disinformation in the Context of Scientific Uncertainty – The Ongoing Case of the COVID-19 'Infodemic', Disinformation and Digital Media as a Challenge for Democracy, pp 367-387, Intersentia, Vol. 6, June 2020

Aleksandra Kuczerawy, Fighting online disinformation: did the EU Code of Practice forget about freedom of expression?, Disinformation and Digital Media as a Challenge for Democracy, European Integration and Democracy Series, pp. 291 – 308, Intersentia, Vol. 6, June 2020; https://doi.org/10.1017/9781839700422.017

David Leslie, Christopher Burr, Mhairi Aitken, Josh Cowls, Mike Katell, Morgan Briggs, Artificial Intelligence, Human Rights, Democracy, and the Rule of Law, The Alan Turing Institute, March 2021; https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

Emma Llansó, Joris van Hoboken, Jaron Harambam, Artificial Intelligence, Content Moderation, and Freedom of Expression, Transatlantic Working Group, February 2020; https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Emma Llansó, No amount of "AI" in content moderation will solve filtering's prior-restraint problem, Big Data & Society, Volume 7, Issue 1, April 2020; https://doi.org/10.1177%2F2053951720920686

Mark MacCarthy, Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry, Transatlantic Working Group, February 2020; http://dx.doi.org/10.2139/ssrn.3615726

Claudio Lombardi, The Illusion of a "Marketplace of Ideas" and the Right to Truth, American Affairs Journal, Vol. III, No. 1, February 2019; https://americanaffairsjournal.org/2019/02/the-illusion-of-a-marketplace-of-ideas-and-the-right-to-truth

Manila Principles, May 2015; https://manilaprinciples.org

Nathalie Maréchal, Targeted Advertising Is Ruining the Internet and Breaking the World, VICE: Motherboard, November 2018, https://www.vice.com/en/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world

Nathalie Maréchal, Rebecca MacKinnon and Jessica Dheere, Getting to the Source of Infodemics: It's the Business Model, Ranking Digital Rights, New America, May 2020; https://www.newamerica.org/oti/reports/getting-to-the-source-of-infodemics-its-the-business-model

Chris Marsden, Trisha Meyer, European Parliament, Study on regulating disinformation with artificial intelligence, European Parliamentary Research Service, Scientific Foresight Unit, March 2019; https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf

Mike Masnick, Impossibility Theorem; Content Moderation at Scale is Impossible To Do Well, Tech Dirt, November 2019; https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml

Eleonora Maria Mazzoli and Damian Tambini, Prioritisation Uncovered, The Discoverability of Public Interest Content Online, Council of Europe, DGI(2020)19, November 2020, https://rm.coe.int/publication-content-prioritisation-report/1680a07a57

Lorna McGregor, Daragh Murray, Vivian Ng, international Human Rights Law as a Framework for Algorithmic Accountability, Cambridge University Press, International and Comparative Law Quarterly, April 2019; http://dx.doi.org/10.1017/S0020589319000046

Meedan, 2020 Misinfodemic Report: COVID-19 in Emerging Economies; https://meedan.com/reports/2020-misinfodemic-report-covid-19-in-emerging-economies

Marko Milanovic, Michael N. Schmitt, Cyber Attacks and Cyber (Mis)information Operations During a Pandemic, Journal of National Security Law & Policy, Vol. 11, p. 247, May 2020; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3612019

MIT Technology Review, Zeynep Tufekci, How social media took us from Tahrir Square to Donald Trump, August 2018; https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump

Judith Möller, Damian Trilling, Natali Helberger, Brams van Es, Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity, Information, Communication & Society, Volume 21, Item 4, pp 1-19, March 2018; http://dx.doi.org/10.1080/1369118X.2018.1444076

Yifat Nahmias, Maayan Perel, The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations, Harvard Journal on Legislation, Vol. 58, no. 1, pp. 146-194, February 2021; https://harvardjol.com/wp-content/uploads/sites/17/2021/02/105_Nahmias.pdf

NBC, Brandy Zadrozny and Ben Collins, As vaccine mandates spread, protests follow – some spurred by nurses, August 2021; https://www.nbcnews.com/tech/social-media/vaccine-mandates-spread-protests-follow-spurred-nurses-rcna1654

Christina Nemr and William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age, Park Advisors, March 2019; https://www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf

Norwegian Consumer Council, Time to Ban Surveillance-Based Advertising: The case against commercial surveillance online, June 2021; https://www.forbrukerradet.no/wp-content/uploads/2021/06/20210622-final-report-time-to-ban-surveillance-based-advertising.pdf

Manfred Nowak, UN Covenant on Civil and Political Rights: CCPR Commentary, 2nd Edition, 2005

Fernando Nuñez, Disinformation Legislation and Freedom of Expression, UC Irvine Law Review, Vol. 10, Issue 2, March 2020; https://scholarship.law.uci.edu/ucilr/vol10/iss2/10

NYU Ad Observatory, https://adobservatory.org/missed-ads

Ofcom, Cambridge Consultants, Use of AI in Online Content Moderation, July 2019; https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

Online Censorship, How to appeal; https://onlinecensorship.org/resources/how-to-appeal

Openai, OpenAI API, June 2020; https://openai.com/blog/openai-api

OSCE Representative on Freedom of the Media, Non-paper on the Impact of Artificial Intelligence on Freedom of Expression, March 2020; https://www.osce.org/files/f/documents/b/a/447829.pdf

OSCE Representative on Freedom of the Media, #SAIFE – Spotlight on Artificial Intelligence and Freedom of Expression, July 2020; https://www.osce.org/files/f/documents/9/f/456319_0.pdf

OSCE Representative on Freedom of the Media, SAIFE Policy Manual, December 2021; https://www.osce.org/files/f/documents/8/f/510332_0.pdf

OSCE Representative on Freedom of the Media, Special Report on Handling of the Media During Public Assemblies, October 2020; https://www.osce.org/files/f/documents/2/f/467892.pdf

OSCE Representative on Freedom of the Media, Report on the expert roundtable: International law and policy on disinformation in the context of freedom of the media, May 2021; https://www.osce.org/files/f/documents/f/9/488884_1.pdf

OSCE Representative on Freedom of the Media, Policy Brief on Disinformation and media self-regulation, June 2021; https://www.osce.org/files/f/documents/d/7/490331_1.pdf

OSCE Representative on Freedom of the Media, Policy Brief on Artificial Intelligence and Disinformation as a Multilateral Policy Challenge, December 2021; https://www.osce.org/files/f/documents/d/0/506702.pdf

Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information, Cambridge, London, Harvard University Press, January 2015

Nazia Parveen and Jim Waterson, UK phone masts attacked amid 5G-coronavirus conspiracy theory, April 2020, The Guardian; https://www.theguardian.com/uk-news/2020/apr/04/uk-phone-masts-attacked-amid-5g-coronavirus-conspiracy-theory

Sara Pluviano, Sergio Della Sala, Caroline Watt, The effects of source expertise and trustworthiness on recollection: the case of vaccine misinformation, Cognitive Processing, Vol. 21, pp. 321-330, April 2020; https://link.springer.com/article/10.1007%2Fs10339-020-00974-8

POLITICO, Mark Scott, Facebook's attempt to ban academics runs into trouble, August 2021; https://www.politico.eu/article/facebook-nyu-laura-edelson-political-ads

POLITICO, Mark Scott, Facebook's private groups are abuzz with coronavirus fake news, March 2020; https://www.politico.eu/article/facebook-misinformation-fake-news-coronavirus-covid19

POLITICO, Mark Scott, Campaigners wanted more transparency. Facebook threatened to sue, August 2020; https://www.politico.eu/article/facebook-transparency-privacy-matthias-spielkamp-algorithm-watch/?utm_source=POLITICO.EU&utm_campaign=281c8a7bd3-EMAIL_CAMPAIGN_2021_08_19_10_15&utm_medium=email&utm_term=0_10959edeb5-281c8a7bd3-190567583

POLITICO, Clothilde Goujard, Facebook faces wrath of EU lawmakers working on online content rules, October 2021; https://www.politico.eu/article/facebook-faces-wrath-of-eu-lawmakers-working-on-online-content-rules

Jennifer L. Pomeranz, Aaron R. Schwid, Governmental actions to address COVID-19 misinformation, Journal of Public Health Policy, Vol. 42, Issue 2, pp 201-210, January 2021; https://doi.org/10.1057/s41271-020-00270-x

Julie Posetti and Kalina Bontcheva, UNESCO, Disinfodemic: Deciphering COVID-19 disinformation, Policy brief 1, April 2020; https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf

Julie Posetti and Kalina Bontcheva, UNESCO, Disinfodemic: Dissecting responses to COVID-19 disinformation, Policy brief 2, May 2020; https://en.unesco.org/sites/default/files/disinfodemic_dissecting_responses_covid19_disinformation.pdf

Protocol, Issie Lapowsky and Ben Brody, Lawmakers want to force Big Tech to give researchers more data, May 2021; https://www.protocol.com/policy/social-media-data-act

Raxona Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation, Social Media + Society, Volume 6, Issue 3, July-September 2020; https://doi.org/10.1177%2F2056305120948190

Ranking Digital Ranking, 2020 Accountability Index, April 2021; https://rankingdigitalrights.org/index2020

Andrei Richter, Fake News and Freedom of the Media, Journal of International Media & Entertainment Law, Volume 8, Number 1, pp 1-34, 2019; https://www.swlaw.edu/curriculum/law-review-journals/journal-international-media-entertainment-law

Ann Cathrin Riedel, Behind closed curtains, Disinformation on messenger services, Friedrich Naumann Foundation For Freedom, August 2020; https://www.freiheit.org/iaf/behind-closed-curtains-disinformation-messenger-services

Bianca Residorf, Grant Blank, Johannes Bauer, Shelia Cotten, Craig Robertson, Megan Knittel, Information Seeking Patterns and COVID-19 in the United States, Journal of Quantitative Description: Digital Media, Volume 1, April 2021; https://doi.org/10.51685/jqd.2021.003

Reuters, J. Scott Brennen, Felix Simon, Philip N. Howard, Rasmus Kleis Nielsen, Types, sources, and claims of COVID-19 misinformation, April 2020; https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Reuters Institute, University of Oxford Digital News Report, March 2021; https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Sarah T. Roberts, Digital detritus: 'Error' and the logic of opacity in social media content moderation, First Monday, Vol. 23, No. 3-5, March 2018; https://doi.org/10.5210/fm.v23i3.8283

Barrie Sander, Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, Fordham International Law Journal, Vol. 43, No. 4, pp 939-1006, May 2020; https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3

Santa Clara Principles on Transparency and Accountability in Content Moderation, February 2018; https://santaclaraprinciples.org

Giovanni Sartor, Andrea Loreggia, European Parliament, The impact of algorithms for online content filtering or moderation: "Upload filters", study requested by the JURI committee, September 2020;

https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf

Antonella Sciortino, Fake News and Infodemia at the Time of Covid-19, Direito Publico, Vol. 17, no. 94, pp 35-49, August 2020; https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/4823

Carey Shenkman, Dhanaraj Thakur, Emma Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, May 2021; https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf

Sarah Shugars, Nicholas Beauchamps, Why Keep Arguing? Predicting Engagement in Political Conversations Online, SAGE, March 2019; https://doi.org/10.1177/2158244019828850

Sarah Shugars, Adina Gitomer, Stefan McCabe, Ryan J. Gallagher, Kenneth Joseph, Nir Grinberg, Larissa Doroshenko, Brooke Foucault Welles, David Lazar, Pandemics, Protests, and Publics: Demographic Activity and Engagement on Twitter in 2020, Journal of Quantitative Description: Digital Media 1, pp 1-68, April 2021; https://doi.org/10.51685/jqd.2021.002

Jacob Silverman, Facebook is Designed to Spread Covid Misinformation, The Soapbox, 19 July 2021; https://newrepublic.com/article/163002/facebook-designed-spread-covid-misinformation

Matthew Spradling, Jeremy Straub, Jay Strong, Protection from 'Fake News': The Need for Descriptive Factual Labeling for Online Content, Future Internet, Volume 13, 142, May 2021; https://doi.org/10.3390/fi13060142

Eduardo Suárez, Reuters, How fact-checkers are fighting coronavirus misinformation worldwide March 2020; https://reutersinstitute.politics.ox.ac.uk/risj-review/how-fact-checkers-are-fighting-coronavirus-misinformation-worldwide

Nicolas P. Suzor, Lawless: the Secret Rules That Govern Our Digital Lives, Cambridge University Press, 2019

Nabiha Syed, Real Talk About Fake News: Towards a Better Theory for Platform Governance, The Yale Law Journal, Volume 127, pp 337-357, October 2017; https://www.yalelawjournal.org/forum/real-talk-about-fake-news

Damian Tambini, Media Freedom, Regulation and Trust: A Systemic Approach to Information Disorder, Artificial Intelligence – Intelligent Policies: Challenges and opportunities for media and democracy, Background Paper, Council of Europe, Ministerial Conference, February 2020; https://rm.coe.int/cyprus-2020-new-media/16809a524f

Tech Transparency Project, Google is Paying Creators of Misleading Coronavirus Videos, April 2020; https://www.techtransparencyproject.org/articles/google-paying-creators-of-misleading-coronavirus-videos

Dhanaraj Thakur, DeVan L. Hankerson, Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender; Center for Democracy & Technology, February 2021; https://cdt.org/wp-content/uploads/2021/02/2021-02-10-CDT-Research-Report-on-Disinfo-Race-and-Gender-FINAL.pdf

Dhanaraj Thakur, Guest Post, Mallory Knodel, Emma Llansó, Greg Nojeim and Caitlin Vogus, Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems, Center for Democracy & Technology, August 2021; https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems

The Guardian, The Cambridge Analytica Files, March 2018; https://www.theguardian.com/news/series/cambridge-analytica-files

The Verge, Facebook was marking legitimate news articles about the coronavirus as spam due to a software bug, March 2020; https://www.theverge.com/2020/3/17/21184445/facebook-marking-coronavirus-posts-spam-misinformation-covid-19

The Virality Project, COVID-19 vaccine policies, February 2021; https://www.viralityproject.org/policy-analysis/evaluating-covid-19-vaccine-policies-on-social-media-platforms

The Wall Street Journal, The Facebook Files, September 2021; https://www.wsj.com/articles/the-facebook-files-11631713039

The Wall Street Journal, The Facebook Files, Part 6: The Whistleblower, October 2021; https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-6-the-whistleblower/b311b3d8-b50a-425f-9eb7-12a9c4278acd?mod=series_facebookfiles

TikTok, COVID-19 policies; https://www.tiktok.com/safety/en-us/covid-19

Toronto Declaration, May 2018; https://www.torontodeclaration.org

Zeynep Tufeki, YouTube, the Great Radicalizer, New York Times Opinion, March 2018; https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Twitter, COVID-19 policies; https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

Twitter Blog, Keith Coleman, Introducing Birdwatch, a community-based approach to misinformation, January 2021; https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

Twitter, An update to the Twitter Transparency Center, July 2021; https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center?utm_source=POLITICO.EU&utm_campaign=a78ce40264-EMAIL_CAMPAIGN_2021_07_15_11_42&utm_medium=email&utm_term=0_10959edeb5-a78ce40264-190567583

Twitter, Guy Rosen on restoring incorrectly removed COVID-19 posts, March 2020; https://twitter.com/guyro/status/1240088303497400320?s=20

Heidi Tworek, Paddy Leerssen, Transatlantic Working Group, An Analysis of Germany's NetzDG Law, April 2019; https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

UK House of Commons, Digital, Culture, Media and Sport Committee, Misinformation in the COVID-19 Infodemic, Secord Report of Session 2019-21, July 2020, United Kingdom of Great Britain and Northern Ireland, Parliament; https://committees.parliament.uk/publications/1954/documents/19089/default

UK House of Lords, Communications and Digital Committee, Free for all? Freedom of expression in the digital age, July 2021; https://committees.parliament.uk/publications/6878/documents/72529/default

UNESCO, Resource Center of Responses to COVID-19, https://en.unesco.org/covid19/communicationinformationresponse/mediasupport

UNESCO, Steering AI and Advanced ICTs for Knowledge Societies: A Rights, Openness, Access, and Multi-stakeholder Perspective, November 2019; https://unesdoc.unesco.org/ark:/48223/pf0000372132.locale=en

UNESCO, Journalism, press freedom and COVID-19, World Trends in Freedom of Expression and Media Development, May 2020; https://unesdoc.unesco.org/ark:/48223/pf0000373573?posInSet=1&queryId=0216815c-9a38-457c-8e20-b224c31b03e5

UNESCO, The Right to Information in Times of Crisis: Access to Information – Saving Lives, Building Trust, Bringing Hope!, World Trends in Freedom of Expression and Media Development, June 2020; https://unesdoc.unesco.org/ark:/48223/pf0000374369

UNESCO, Letting the Sun Shine In: Transparency and Accountability in the Digital Age, World Trends in Freedom of Expression and Media Development, May 2021; https://unesdoc.unesco.org/ark:/48223/pf0000377231

United Nations General Assembly, Seventy-third session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on "artificial intelligence", David Kaye, A/73/348, August 2018; https://undocs.org/pdf?symbol=en/A/73/348

United Nations General Assembly, Seventy-fourth session, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and

expression on "hate speech", David Kaye, A/74/486, October 2019; https://undocs.org/en/A/74/486

United Nations Interregional Crime and Justice Research Institute, Stop the virus of disinformation. The risk of malicious use of social media during COVID-19 and the technology options to fight it, November 2020; http://www.unicri.it/sites/default/files/2020-11/SM%20misuse.pdf

United Nations Office of the High Commissioner for Human Rights, The European Union and International Human Rights Law, https://europe.ohchr.org/Documents/Publications/EU_and_International_Law.pdf

United Nations, Sustainable Development Goals – Transforming our world: the 2030 Agenda for Sustainable Development, https://sdgs.un.org/2030agenda

Joris van Hoboken, Ronan Ó Fathaigh, Regulating Disinformation in Europe: Implications for Speech and Privacy, UC Irvine Journal of International, Transnational, and Comparative Law, Volume 6 Symposium: The Transnational Legal Ordering of Privacy and Speech, Article 3, May 2021; https://scholarship.law.uci.edu/ucijil/vol6/iss1/3

Soroush Vosoughi, Deb Roy, Sinan Aral, The spread of true and false news online, Science, Vol. 359, Issue 6380, pp. 1146-1151, MIT study, March 2018, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf

Ben Wagner; Ethics As An Escape From Regulation. From "Ethics-Washing" to Ethics-Shopping?, Amsterdam University Press, December 2019; https://doi.org/10.1515/9789048550180-016

Ben Wagner, Johanne Kübler, Eliška Pírková, Rita Gsenger, Carolina Ferro, Reimagining Content Moderation and Safeguarding Fundamental Rights: A Study on Community-Led Platforms, European Parliament Greens/EFA study, May 2021; https://enabling-digital.eu/wp-content/uploads/2021/07/Alternative-content_web.pdf

Claire Wardle, Hossein Derakhshan, Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe report, DGI(2017)09, September 2017; https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Wikipedia, List of Google April Fools' Day jokes; https://en.wikipedia.org/wiki/List_of_Google_April_Fools%27_Day_jokes#2020%E2%80%9321:_cancellation

Wikiquote, Misquotations; https://en.wikiquote.org/wiki/Misquotations

Thomas Wood, Ethan Porter, The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence, Political Behavior, Volume 41, Issue 1, pp 135-163, March 2019; https://doi.org/10.1007/s11109-018-9443-y

World Health Organization (WHO), Timeline: WHO's COVID-19 response; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline

World Health Organization (WHO), 1st WHO Infodemiology Conference, June 2020; https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference

World Wide Web Consortium (W3C), World Wide Web Foundation, Internet live stats; https://www.internetlivestats.com

Jillian C. York, Corynne McSherry, Content Moderation is Broken, Let Us Count the Ways, Electronic Frontier Foundation (EFF), April 2019; https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways

Jonathan Zittrain, The Inexorable Push for Infrastructure Moderation from the it's-coming-whether-we-like-it-or-not dept, Tech Policy Greenhouse by Techdirt, September 2021; https://www.techdirt.com/articles/20210924/12012347622/inexorable-push-infrastructure-moderation.shtml

Tara Zimmerman, Introducing the Concept of Social Noise, University of North Texas, October 2020; http://hdl.handle.net/2142/108848

Ethan Zuckerman, I read Facebook's Widely Viewed Content Report. It's really strange. August 2021; https://ethanzuckerman.com/2021/08/18/facebooks-new-transparency-report-is-really-strange

Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, New York: Public Affairs, 2019

7.2. Jurisprudence

Abrahams v United States, Case 250 U.S. 616 at 630 (1919), judgment, 10 December 1919, https://supreme.justia.com/cases/federal/us/250/616

European Court of Human Rights, Handyside v. the United Kingdom, application no. 5493/72, judgment, 7 December 1976; http://hudoc.echr.coe.int/eng?i=001-57499

European Court of Human Rights, Lingens v. Austria, application no. 9815/82, judgment, 8 July 1986; http://hudoc.echr.coe.int/eng?i=001-57523

European Court of Human Rights, Autronic AG v. Switzerland, application no. 12726/87, judgment, 22 May 1990; http://hudoc.echr.coe.int/eng?i=001-57630

European Court of Human Rights, Refah Partisi (the Welfare Party) and Others v. Turkey, application no. 41340/98, 41342/98, 41343/98 and 41344/98, judgment, 13 February 2003; http://hudoc.echr.coe.int/eng?i=001-60936

European Court of Human Rights, Salov v. Ukraine, application no. 65518/01, judgment, 6 September 2005; http://hudoc.echr.coe.int/eng?i=001-70096

European Court of Human Rights, Kwiecien v. Poland, application no. 51744/99, judgment, 9 January 2007; http://hudoc.echr.coe.int/eng?i=001-115705

European Court of Human Rights, Ahmet Yildirim v Turkey, application no. 3111/10, judgement, 18 December 2012; http://hudoc.echr.coe.int/eng?i=001-115705

European Court of Human Rights, Delfi AS v. Estonia, judgment, 16 June 2015; http://hudoc.echr.coe.int/eng?i=001-155105

European Court of Justice, Roman Zakharov v Russia, application no. 47143/06, judgment, 4 December 2015; http://hudoc.echr.coe.int/eng?i=001-159324

European Court of Human Rights, Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary, application no. 22947/13, judgment, 2 Feburary 2016; http://hudoc.echr.coe.int/eng?i=001-160314

European Court of Human Rights, Huseynova v. Azerbaijan, application no. 10653/10, judgment, 13 April 2017; http://hudoc.echr.coe.int/eng?i=001-172661

European Court of Human Rights, Catt v. United Kingdom, application no. 43514/15, judgment, 24 January 2019; http://hudoc.echr.coe.int/eng?i=001-189424

European Court of Justice, Eva Glawischnig-Piesczek v. Facebook Ireland Limited, C-18/18, judgment, 3 October 2019; https://curia.europa.eu/juris/liste.jsf?num=C-18/18