Submission to UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression report on disinformation

February 15, 2021

Dr MacKenzie F. Common, Reuters Institute for the Study of Journalism, University of Oxford

Professor Rasmus Kleis Nielsen, Reuters Institute for the Study of Journalism, University of Oxford

Introduction

Disinformation, here understood in line with the UN Special Rapporteur's call for input as "false information that is created and spread, deliberately or otherwise, to harm people, institutions and interests", represents a range of serious issues that can be part of distortions of electoral processes, incitements to violence, and can fuel dangerous conspiracy theories. The coronavirus pandemic has further underlined how disinformation can also represent a risk to personal and public health.

More broadly, behaviours and forms of expression discussed under the heading disinformation often overlap with misinformation (false or misleading information, spread without intent to harm) and malinformation (information that is not false, but strategically used with intent to harm) and with wider discussions under the imprecise and misleading but frequently used term "fake news" (a term that is unfortunately often used to refer to material that is neither fake nor news).[1]

Beyond the actual harm and risk of harm posed by various kinds of disinformation and in some instances by broader kinds of problematic information, the lack of conceptual clarity in defining the problem, and frequent lack of substantial agreement what exact kinds of behaviour and content are problematic, are in our view key parts of the problems we face in addressing these problems. If we do not know, or do not agree, what disinformation is, it will be hard to address it in effective and proportional ways. The lack of clarity and lack of agreement undermine our ability to precisely address specific problems, and undermine the legitimacy of interventions against kinds of content some might consider to be disinformation but others regard as legitimate speech. It also underlines the inherently political nature of determining what does and does not constitute disinformation.

In this short submission, we (1) summarize a set of empirical research findings on disinformation, misinformation, and malinformation that we hope will be helpful in developing measures responding to disinformation that effectively protect free expression, including independent journalism and news media, (2) identify certain risks to free expression that we believe are illustrated by some recent steps by a number of governments and platform companies, and (3) offer some recommendations that we hope will be useful as part of the wider discussion.

Our submission is not meant to be exhaustive and instead focuses on specific areas that we research. We mention other relevant work in passing but will not try to summarize an extensive research area in this short submission.

[1] Wardle, Claire, and Hossein Derakhshan. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Report to the Council of Europe. https://shorensteincenter. org/information-disorder-framework-for-research-and-policymaking.

Empirical research on disinformation, misinformation, and malinformation

Disinformation, misinformation, and malinformation is clearly widespread, especially online. Like much online communication of a more benign or ambivalent character, problematic information is often distributed via platforms, especially popular social media platforms such as Facebook, popular video sharing sites such as YouTube (owned by Google), and popular messaging applications such as WhatsApp (owned by Facebook).[2] Problematic information can also be accessed via search engines such as Google Search and various much smaller competitors (Bing, etc.). In addition to sometimes being surfaced by algorithmic ranking systems controlled by platform companies, it is also sometimes monetized by programmatic advertising services offered by the same companies. Smaller platforms such as Twitter, Snapchat, TikTok, Telegram and various others can play an important role too, especially for specific forms of problematic information or communities who have been de-platformed elsewhere. There is a real risk that the technical and commercial systems run by some of these companies can exacerbate some disinformation problems due to the way in which they incentivize actors (whether political, after profit, or with other motivations) through ranking decisions and through the flow of attention, advertising revenues and other valuable, scare resources.

While clearly widespread and problematic, in countries where systematic academic research exists, problematic information is still a small subset of all the information circulating. This takes nothing away from the seriousness of some of the problems we face, or the urgency to address them, but an evidence-based understanding of scale and scope is a necessary part of any attempt to assess what measures against disinformation are necessary and proportional.

Attempts at measurement are complicated by the above-mentioned lack of definitional clarity and substantial agreement, but some important examples include peer-reviewed academic work finding (1) that identified "fake news" constitutes about 0.15% of American's daily media diet, (2) that, in the run-up to the 2016 election, three in four Americans did not visit an identified “fake news” website, but one in four did so at least once, with the bulk of visits heavily concentrated among the 10% of people with the most conservative online information diets, (3) that a small group of heavy Internet users are the main users of identified “fake news” websites, and that these users are often even more engaged, and far more loyal, users of established news media's websites, and (4) that exposure to “fake news” may have limited effects aside from increasing beliefs in false claims (and thus will not necessarily influence other forms of behaviour, whether in terms of e.g. voting or vaccine uptake).[3]

Despite growing evidence that exposure to and engagement with disinformation narrowly defined on the basis of identified problematic domains is a very small part of most people's media use,

[2] Marwick, Alice E. 2018. “Why Do People Share Fake News? A Sociotechnical Model of Media Effects.” Georgetown Law Technology Review 2 (2): 474–512.

[3] See (1) Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. “Evaluating the Fake News Problem at the Scale of the Information Ecosystem.” Science Advances 6 (14): eaay3539. https://doi.org/10.1126/sciadv.aay3539, (2) Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020. “Exposure to Untrustworthy Websites in the 2016 US Election.” Nature Human Behaviour 4 (5): 472–80. https://doi.org/10.1038/s41562-020-0833-x, (3) Nelson, Jacob L., and Harsh Taneja. 2018. “The Small, Disloyal Fake News Audience: The Role of Audience Availability in Fake News Consumption:” New Media & Society, February. https://doi.org/10.1177/1461444818758715, and (4) Guess, Andrew M., Dominique Lockett, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2020. “‘Fake News’ May Have Limited Effects beyond Increasing Beliefs in False Claims.” Harvard Kennedy School Misinformation Review 1 (1). https://doi.org/10.37016/mr-2020-004.

concentrated among partisans actively seeking it out, and often primarily consumed by people who consume far more news from established outlets, survey research suggests very widespread concern over disinformation, especially online. One survey conducted in 2020 asked respondents across 40 markets whether, thinking about online news, they were concerned about what is real and what is fake on the internet – 56% of respondents across these markets were worried about this, ranging from a low 32% in the Netherlands to a high 84% in Brazil.[4]

This widespread concern over the credibility, quality, and veracity of information online underlines the need to address problems of disinformation and broader kinds of problematic information, but it is also important to recognize what the public see as the main drivers of these problems, both in terms of the platforms were they might encounter it and the actors people see as spreading it. When asked about platforms, public concern is clearly focused on social media, especially Facebook. 40% of respondents across the 40 markets identify social media as the platform where they are most concerned about coming across false or misleading information. When asked about actors, 40% of respondents identify the government, politicians or political parties in their own country as the source they are most concerned about false or misleading information from, followed by 14% who identify activists, 13% journalists, 13% ordinary people, and 10% foreign governments.[5]

It is clear that from the point of view of the public, disinformation is to a large extent a problem associated with the behaviour of politicians and other domestic actors, especially on social media, and not more narrowly a problem of false information or actors with more unambiguously ill intent. This is in line with earlier research which has identified that what people see as stories where facts are twisted to push an agenda (political propaganda) and what people see as examples of poor journalism (superficial, sensationalist, inaccurate content) are among the types of potential misinformation that people in most countries say they most frequently encounter and that the highest number of people say they are concerned about.[6] This is aligned with research showing how prominent public figures including elected officials account for a large share of social media engagement with misinformation and other work suggesting that mainstream news media sometimes risk playing "a significant and important role in the dissemination of fake news".[7]

The combination of (a) lack of conceptual clarity and substantial agreement on what exactly constitutes disinformation, (b) empirical research suggesting that, at least in the countries where such work has been done, identifiable "fake news" makes up a very small part of people's overall media use, and (c) very widespread public concern over problematic information online, concern focused not on demonstrably false information spread to harm (disinformation narrowly defined), but on much broader categories of political propaganda and poor journalism together represents one of the key challenges in addressing the real harm and risks posed by disinformation.

[4] Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andı, and Rasmus Kleis Nielsen. 2020. "Reuters Institute Digital News Report 2020." Oxford: Reuters Institute for the Study of Journalism. http://www.digitalnewsreport.org/.

[5] Newman et al 2020.

[6] Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David A. L Levy, and Rasmus Kleis Nielsen. 2018. "Reuters Institute Digital News Report 2018." Oxford: Reuters Institute for the Study of Journalism. http://www.digitalnewsreport.org/.

[7] Brennen, J. Scott, Felix S. Simon, Philip N. Howard, and Rasmus Kleis Nielsen. 2020. "Types, Sources, and Claims of COVID-19 Misinformation." Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation, Tsfati, Yariv, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren. 2020. "Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis." Annals of the International Communication Association 44 (2): 157–73. https://doi.org/10.1080/23808985.2020.1759443.

They risk creating a situation where measures meant, at least nominally, to address very specific problems of narrowly defined types of disinformation, for political reasons or in response to much wider public concern, end up restricting much broader terrains of information that may be problematic, but are often neither demonstrably harmful nor demonstrably false. Furthermore, they would expand attempts to counter disinformation to forms of speech that would normally often be protected under the human right to impart information and ideas, which, as the United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information pointed out in their 2017 joint statement, "is not limited to "correct" statements, that the right also protects information and ideas that may shock, offend and disturb".[8]

Risks to free expression illustrated by some recent steps by some governments and platform companies

Platform companies have their own content moderation policies. In many cases, these have historically been very permissive on political issues, in line with the First Amendment tradition of the US where many of these companies were founded and are headquartered, even as they have been more restrictive on some specific issues (such as nudity) in ways that reflect a mix of commercial and cultural considerations. How these policies are implemented in practice varies, and, like other aspects of how platforms operate, sometimes seem to disadvantage already historically marginalized and disadvantaged communities.[9] At least on paper, the policies are generally meant to apply equally to all users everywhere. It is important to note that, while these content moderation policies are often significantly more restrictive than US laws regulating free speech, they can be more permissive than local laws across the world that are often more restrictive than those found in the US.

Platform companies' content moderation policies and their practical enforcement has already led to a number of instances where important forms of free speech and the work of journalists and independent news media have been restricted in problematic ways – ranging from YouTube removing content documenting the civil war in Syria to Facebook removing articles accompanied by the iconic photo of Phan Thị Kim Phúc running naked after being severely burned on her back by a South Vietnamese napalm attack.[10]

Especially since 2016, and even more so in the course of the coronavirus pandemic, many platform companies' content moderation policies have been revised and expanded in part to cover a wider range of problematic information, including various kinds of disinformation, misinformation, and malinformation. Companies have expanded their policies and in some cases in enforcement to various degrees even as disinformation problems have also continued to evolve.

While in some ways welcome, these expansions also come with the risk of restricting legitimate speech, and as they are often enforced inconsistently, with little transparency, and no independent

[8] https://www.osce.org/files/f/documents/6/8/302796.pdf

[9] Caplan, Robyn. 2018. "Content or Context Moderation?" Data & Society. https://datasociety.net/library/content-or-context-moderation/, Noble, Safiya Umoja. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. New York: University Press.

[10] https://www.eff.org/deeplinks/2017/12/seven-times-2017-journalists-were-censored

oversight or due process. There are also real risks associated with the enforcement of these policies, whether through artificial intelligence systems, human content moderation or, more commonly, some combination. While automation can be used to scale up content moderation to deal with things at great scale and great pace, the very real limitations of necessarily imperfect technologies combined with the inherently political nature of decisions over what constitutes disinformation means there are serious practical and principled limitations to how useful artificial intelligence will be in dealing with disinformation.

Increasingly, however, governments take an active and direct role in content moderation online, issues that were in practice left more or less to private companies in much of the world. There are clearly many instances in which governments taking a more active role, on the basis of clear and precise legislation, and in ensuring independent oversight, transparency, and due process, is entirely appropriate. But there is also a risk that some governments will pursue responses to disinformation that – irrespective of whether they in fact help address specific problems of narrowly defined types of disinformation – risk restricting free speech.[11]

Governments may, for example, pass laws that define disinformation as including, among other things, content that is critical of the government or counters government messaging. A controversial law in Pakistan provides no definition of fake news and states that content should be labelled as false if the Pakistani regulatory authority says it is false.[12] Similarly, Vietnam's Law on Cybersecurity has a broad prohibition on disinformation (although it is only labelled as "conduct which is strictly prohibited") which includes "distorting history, denying revolutionary achievements, destroying the national solidarity block" and "providing false information, causing confusion amongst the citizens," and "cheating or tricking, manipulating, training or drilling people to oppose the State of the Socialist Republic of Vietnam."[13] Disinformation laws that are too broad and vague or pose a risk to human rights can, therefore, like similarly broad and vague laws already on the books, risk chilling legitimate speech and can be used selectively or indiscriminately by governments to encourage or require private companies to police speech in ways that can harm free expression and limit public debate.

Governments can, among other things, use such laws, or already existing legislation with broad and vague prohibitions on e.g. blasphemy, Lèse-majesté, or sedition, to co-opt platforms into geo-blocking content (commonly known as Country-Withheld-Content applications) or removing it entirely. Governments can then ensure that regardless of whether they initiate prosecutions (if the user can even be identified and is located within that jurisdiction), the content will not be available in the given jurisdiction. Country-Withheld-Content (CWC) actions require that countries notify platforms of content that is in violation of local law and request that the content be made inaccessible in that jurisdiction.[14]

[11] Lim, Gabrielle. 2020. "Securitize/Counter-Securitize The Life and Death of Malaysia's Anti-Fake News Act." New York: Data and Society. https://datasociety.net/library/securitize-counter-securitize/ .

[12] Rule 5(f), *Citizens Protection (Against Online Harm) Rules,* 2020, Pakistan.

[13] Article 8(1)a-d, *Law on Cybersecurity,* 2018, Vietnam. See also, "Let us Breathe: Censorship and Criminalization of Online Expression in Viet Nam." *Amnesty International* (2020) https://www.amnestyusa.org/wp-content/uploads/2020/11/LET-US-BREATHE-ASA41_3243_2020.pdf 17.

[14] Twitter, for example, states: "Many countries, including the United States, have laws that may apply to Tweets and/or Twitter account content. In our continuing effort to make our services available to people everywhere, if we receive a valid and properly scoped request from an authorized entity, it may be necessary to withhold access to certain content in a particular country from time to time. Such withholdings will be limited to the specific jurisdiction that has issued the valid legal demand or where the content has been found to violate local law(s)." Twitter Help Centre, "About Country Withheld Content" https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country Accessed 11 February 2021.

While platforms, like other private companies, generally say they will abide by local laws, it should be noted that they do not typically approve all requests for geo-blocking or requests for removal and that they have sometimes pushed back against government attempts to get them to restrict specific forms of speech. However, their exact decision-making methods remain opaque and difficult for outsiders to predict. Despite the existence of databases like Lumen, there is limited transparency not only on content moderation broadly, but also content moderation based on government requests.[15]

Governments have different ways in which they can exert pressure on platforms to restrict more content in their jurisdiction. In April 2020, Facebook announced that it had agreed to "significantly increase" compliance with requests from the Vietnamese government to censor "anti-state" content in Vietnam.[16] Facebook's decision came after the Vietnamese government's decision to take Facebook's local servers offline for seven weeks, making the platform slow and inoperable in Vietnam.[17] Since the announcement, *Amnesty International* reports, there has been a 983% increase in Facebook restricting content within Vietnam based on local law and Facebook has complied with 95% of the government's content requests.[18] In late 2020 and early 2021, YouTube and TikTok complied with Turkey's recently amended internet law and appointed a local representative, making it much more susceptible to content removal and take-down requests by the Turkish authorities.[19] At the time of writing, Facebook, Pinterest, and Twitter were among platform companies still resisting the requirement, facing fines and potentially a ban on Turkish companies advertising on the platforms as a consequence. Developments in these countries illustrate some of the leverage governments have over platform companies, including holding individual local representatives responsible as well as the ability to lever fines, block access, and other ways of impacting the companies' bottom line.

In most cases, to influence platforms' content moderation directly, governments must request content be geo-blocked by reaching out to platforms and providing them with information about the offending content and relevant laws. This process takes time and is usually done by government employees (for example civil servants in the executive branch or by law enforcement) or a designated government regulator. Because of the limitations built into this process, which can be time consuming, a significant proportion of content that might violate local laws (including against disinformation) but is not identified or deemed to have violated a given platform's own rules, will often remain accessible.

This can be a serious issue for governments and can lead to uncertainty and inconsistency, even as it may also in practice limit attempts at restricting content (similar to the idea of "practical obscurity" - which relies on the difficulty of accessing information - as an early method of protecting privacy). Inspired by the German Network Enforcement Act of 2017, countries are beginning to try to close what could be seen as an enforcement gap through new types of legislation. The Network Enforcement Act obliges platforms that have over two million users in Germany to provide a system of handling reports about illegal content. The main platforms all have systems that allowed users to flag content that is in violation of the platform's rules but now, they also allow German users to flag content that they believe is illegal under German law. Instead of relying on government regulators to

[15] https://lumendatabase.org/

[16] James Pearson, "Exclusive: Facebook agreed to censor posts after Vietnam slowed traffic-sources." *Reuters*, 2020. https://www.reuters.com/article/us-vietnam-facebook-exclusive-idUSKCN2232JX

[17] Pearson, "Exclusive: Facebook agreed to censor posts after Vietnam slowed traffic-sources."

[18] YouTube, which was also a target of Vietnamese government pressure, complied with 90% of all requests. Let us Breathe: Censorship and Criminalization of Online Expression in Viet Nam. 6, 23.

[19] https://www.hrw.org/news/2020/12/19/turkey-youtube-precedent-threatens-free-expression

actively notify platforms, this Act was able to transform every social media user in Germany into an agent of the law, significantly scaling up the volume of notices. Platform then has 24 hours to address "manifestly illegal content" or 7 days to investigate and make a decision or else they will be fined, significantly accelerating the pace of moderation. The impact of the Network Enforcement Act in Germany specifically may have been less severe than many human rights groups feared when it was introduced.[20] But risks remains, including that countries with poor human rights records will model disinformation laws on the German Network Enforcement Act, thus exponentially enhancing their enforcement capabilities and leading to mass flagging and restrictions, potentially further curtailing free expression.

These developments illustrate the fact that some governments are in different ways increasingly encouraging platform companies to undertake regulatory and police functions that are traditionally considered a matter of public law. Functions are delegated to platforms by government regulation on the one hand and, on the other hand, platform companies sometimes seek to assume such functions, perhaps in part to reduce their liability and the risk of more regulation.[21] Unless attempts to counter disinformation in these ways are clearly prescribed by law, justified as necessary to protect specific legitimate values and identities, proportional to the problems in question, and provide transparency and due process, such measures confer significant discretion on platforms, who will often have incentives to err on the side of caution. This is a highly problematic situation from the standpoint of the protection of individual rights, and carries significant risks for free expression including independent news media.

Recommendations

As said at the outset we do not provide exhaustive overview of research, issues, or recommendations, but will close with highlighting some evidence-based practical responses and a few observations on how issues and risks arising from some possible responses to disinformation could be handled in ways that will protect free expression and independent news media.

<u>Practical responses:</u> It is important to recognize that empirical research identifies a number of practical responses to different kinds of disinformation that have been successful at reducing its effect directly or indirectly, reducing its spread, and increasing societal resilience to disinformation problems without infringing on free expression or other fundamental rights. Research has documented the effect of some specific kinds of fact-checking, content labelling/tagging, and media literacy interventions both directly and indirectly.[22] More broadly, countries with diverse and robust

[20] Heldt, Amelie. 2020. "Germany Is Amending Its Online Speech Act NetzDG" Internet Policy Review. Accessed February 15, 2021. https://policyreview.info/articles/news/germany-amending-its-online-speech-act-netzdg-not-only/1464.

[21] Belli, Luca, David Erdos, Maryant Fernández Pérez, Pedro Augusto P. Francisco, Krzysztof Garstka, Judith Herzog, Krisztina Huszti-Orban, et al. 2017. Platform Regulations: How Platforms Are Regulated and How They Regulate Us. FGV Direito Rio. http://bibliotecadigital.fgv.br/dspace/handle/10438/19402.

[22] Wood, Thomas, and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." Political Behavior 41 (1): 135–63. https://doi.org/10.1007/s11109-018-9443-y. Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. "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 42 (4): 1073–95. https://doi.org/10.1007/s11109-019-09533-0. Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India." Proceedings of the National Academy of Sciences 117 (27): 15536–45. https://doi.org/10.1073/pnas.1920498117.

independent news media seem more resilient to disinformation.[23] Such interventions have a proven track record, and do not restrict free expression or other fundamental rights. Instead of restricting speech, they qualify it. Instead of limiting independent news media, they enable them. Governments wishing to counter disinformation are well-positioned to encourage the implementation of such measures by mandating transparency reports documenting who does (and does not) engage in proven examples of good practice, and by providing direct and indirect funding support for independent fact-checking, media literacy, and news media.[24] Similarly, platform companies could implement and/or support such measures, and share data so that independent researchers could regularly assess and test the efficacy of different responses.

Legal responses: It is particularly important that the same standards of human rights protections be applied to online conduct as are applied to offline conduct, and that the enforcement of legal restrictions on online speech are consistent, transparent, and ensure due process. To ensure that measures to counter disinformation protect free expression, states could commit to approaches where any legal restrictions on speech are clearly and precisely prescribed by law, only introduced where they are necessary to protect other fundamental values, and are proportional to the specific threat at hand. These three cumulative conditions are established in the European Convention on Human Rights as it has evolved through the interpretation given to its texts by the European Court of Human Rights, the European Commission of Human Rights, and the work of the Council of Europe, and the Court's established rules for strict interpretation and insistence that the burden to prove that all three requirements are fulfilled falls on the state help provide extra safeguards.[25]

Platform responses: It is important that platforms align their policies and processes with international human rights principles, and point out when they believe that these may be in potential tension with local laws.[26] As has been made clear in a previous report by the UN Special Rapporteur, "human rights standards, if implemented transparently and consistently with meaningful user and civil society input, provide a framework for holding both States and companies accountable to users across national borders." [27] It is clear that states' international human rights law obligations require that they respect, protect and fulfil human rights locally and that companies, in addition to abiding by the law, should respect human rights. The UN Guiding Principles on Business and Human Rights provides a clear basis for taking a stance when local laws and human rights principles conflict. As they state, "the responsibility to respect human rights is a global standard of expected conduct for all business enterprises wherever they operate. It exists independently of States' abilities and/or willingness to fulfil their own human rights obligations".[28]

Oversight, transparency, and due process: greater transparency in how platforms engage in content moderation broadly, and around disinformation specifically, would be an important step. This has

[23] Humprecht, Edda, Frank Esser, and Peter Van Aelst. 2020. "Resilience to Online Disinformation: A Framework for Cross-National Comparative Research." The International Journal of Press/Politics, January, 1940161219900126. https://doi.org/10.1177/1940161219900126.

[24] https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[25] Bychawska-Siniarska, Dominika. 2017. "Protecting The Right To Freedom Of Expression Under The European Convention On Human Rights." Council of Europe. https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814 .

[26] Kaye, David. 2019. Speech Police: The Global Struggle to Govern the Internet. New York, NY: Columbia Global Reports.

[27] UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Report on Content Regulation to the HRC, UN Doc. A/HRC/38/35, para. 41.

[28] https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

already been raised elsewhere, including in the Santa Clara principles on transparency and accountability in content moderation, and in ideas for genuinely independent social media councils with a broad brief as multi-stakeholder accountability mechanism for content moderation by different platform companies.[29] This includes greater transparency, including on the use of artificial intelligence in content moderation around disinformation. While useful for some specific purposes, automation has already been shown to be imprecise when it comes to identifying e.g. copyrighted content and pornography, and will necessarily be even more so when it comes to inherently political judgements about what exactly constitutes disinformation, and should be used cautiously, with human oversights, and ideally independent scrutiny, when it comes to dealing with different kinds of disinformation. Increasingly, such transparency needs to include documentation of the orders and requests platforms receive from governments, and what steps, if any, platforms take in response. It is particularly important that platforms share more information about their CWC programs as these methods are often employed as a response to local law requests and are difficult to analyse from outside the company. Platforms do frequently share the number of CWC requests and their rate of compliance, but more detailed information is required (even, or especially, in cases where governments may not want disclosure). This is particularly important when countries are passing broad disinformation laws that could generate a large number of questionable CWC requests. It would be troubling if we moved from opaque unilateral and often unaccountable content moderation by private companies to opaque unilateral and often unaccountable content moderation by governments. Sadly, it remains the case that free expression is threatened across the world, and that one of the threats comes from some political actors and even governments, even in historically stable long-standing democracies.[30] It would be unfortunate to replace naiveté about platform companies with naiveté about political actors.

These are some partial possible recommendations for taking different steps that could help address various disinformation problems without jeopardizing free expression and other fundamental rights and ensuring greater consistency, transparency, and accountability. They focus on practical interventions and legal and governments responses that largely sidestep directly addressing what research suggests are some of the most important actors when it comes to the spread of problematic information – (some) domestic politicians and (some) media. This is deliberate. It is entirely possible, in some extreme cases perhaps even appropriate and justified to restrict the most egregious examples of elected officials and self-proclaimed news organizations creating and spreading false information, deliberately or otherwise, that harm people, institutions and interests. But such steps should be taken with the utmost caution, and only where hard evidence suggest they are necessary and proportional to the documented harm of disinformation and they can be taken without jeopardizing free expression more broadly.

Such caution does not preclude using research to estimate the scale and scope of disinformation problems we face and identify evidence-based responses, it does not preclude using international human rights law as a basis for encouraging both governments and platform companies to take additional steps (but only steps compatible with human rights), and it does not preclude steps to increase oversight, transparency, and due process when it comes to online content moderation. This will not provide a one-size-fits-all model, or a single solution – we do not think such a model or such

[29] https://santaclaraprinciples.org/ and https://www.article19.org/resources/social-media-councils-consultation/

[30] Nielsen, Rasmus Kleis, Robert Gorwa, and Madeleine de Cock Buning. 2019. "What Can Be Done? Digital Media Policy Options for Europe (and Beyond)." Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/risj-review/what-can-be-done-digital-media-policy-options-europe-and-beyond.

a solution exists. But it provides us with a way of developing legitimate models and effective solutions to protect ourselves and one another form disinformation without undermining our fundamental rights.

About the authors

MacKenzie Common is a Research Fellow at the Reuters Institute for the Study of Journalism at the University of Oxford. She has a PhD in Law from the London School of Economics (LSE). Her research focuses on the content moderation processes at social media companies and argues that many of their practices are problematic from a human rights law and rule of law standpoint.

Rasmus Kleis Nielsen is Director of the Reuters Institute for the Study of Journalism and Professor of Political Communication at the University of Oxford. He has a PhD in Communications from Columbia University. His work focuses on changes in the news media, on political communication, and the role of digital technologies in both.

About the Reuters Institute for the Study of Journalism

The Reuters Institute for the Study of Journalism is dedicated to exploring the future of journalism worldwide through debate, engagement, and research. It is part of the Department of Politics and International Relations at the University of Oxford. Core funding comes from the Thomson Reuters Foundation, as well as from a wide range of other funders including academic funding bodies, foundations, non-profits, and industry partners.