HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy

February 15, 2021

Ms. Irene Khan
Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
OHCHR-UNOG
8-14 Avenue de la Paix
1211 Geneve 10, Switzerland

Re: UN Special Rapporteur's Annual Thematic Report to be Presented to the Human Rights Council at its 47th Session in June 2021

To the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,

The Technology and Social Change Team submits the following comment in response to the UN Special Rapporteur's call for inputs regarding the upcoming report to the Human Rights Council, to be presented in June 2021. The Technology and Social Change team (TaSC) researches media manipulation and disinformation at scale. TaSC conducts research, develops methods, and facilitates workshops for students, journalists, policy makers, technologists, and civil society organizations on how to detect, document, and debunk media manipulation campaigns that seek to control public conversation, derail democracy, and disrupt society. TaSC is led by sociologist Joan Donovan, PhD, Research Director of Harvard Kennedy School's Shorenstein Center, and a field leading expert in online extremism, media manipulation, and disinformation.

DISINFORMATION AT SCALE THREATENS FREEDOM OF EXPRESSION WORLDWIDE

Comment of Joan Donovan, Emily Dreyfuss, Gabrielle Lim, and Brian Friedberg of The Technology and Social Change Team at the Harvard Shorenstein Center[1]

The human right to freedom of expression includes the right to have *access* to such expression. Increasingly, that access is threatened by social inequalities and the technological systems that hold the world's information. Within the fragmented media ecosystem of the 21st century, opaque algorithms, policies, and enforcement mechanisms determine what information is available to whom. These crucial information distribution systems – from search engines to social media, from messaging apps to legacy news publications – are vulnerable to abuse by people wishing to inject false or misleading information into the ecosystem, to cause harm, or further their own agendas. This process is known as disinformation. In the following comment, we argue that mitigating disinformation is not at odds with the right to freedom of expression. Rather, we demonstrate that mitigating disinformation is essential to safeguarding the human right to freedom of expressions and access to truth.

Based on our research and domain expertise, disinformation violates the right to freedom of expression and the right to information and truth in the following ways:

1. It makes it harder to access timely, relevant, and accurate information
2. It takes advantage of algorithmic amplification to intentionally mislead
3. It silences its target victims through harassment, incitement of fear, and by crowding out their words, opinions, and other forms of expression

We do not dispute that those wishing to spread disinformation have a right to express themselves. However, we point out that the right to freedom of expression does not convey the right to have that disinformation amplified at scale, and that by doing so, may

[1] Authors thank Spring 2021 Harvard Law School Cyberlaw Clinic students Clara Carvahlo e Silva and Melyssa Eigen for their valuable assistance in preparing this comment.

actually lead to self-censorship, oppression, and other harmful effects that are counter to a democratic society.

This is in large part due to the internet's network effect that can accelerate the spread of disinformation and massively increasing the number of people it may reach. Social media, especially, brings with it mechanisms and tactics that allow for large-scale coordinated disinformation campaigns that are often hard to recognize and nearly impossible to mitigate once they have reached millions. The effect of some disinformation campaigns is real world harm, such as hate crimes, violence, harassment, and the perpetuation of discrimination.

To balance the right to express oneself with the right to access the expressions of others, including time-sensitive true and necessary information, we recommend adopting *community-based curation methods* for internet content. We explain that content moderation – the method for handling disinformation most commonly used and advocated for – is reactive and therefore insufficient. But by adopting a proactive curation policy that is grounded in community input, coupled with moderation when necessary, we can create an information ecosystem that promotes truth over sensationalism, accuracy over popularity, and can additionally be subject to more effective oversight.

1. DISINFORMATION HARMS FREEDOM OF EXPRESSION

Disinformation, defined as spreading information that is deliberately false or misleading,[2] directly impedes the right to freedom of expression. The international standard for freedom of expression gives all people the right not only to seek but also to "*receive* . . . information and ideas of all kinds, regardless of frontiers . . . through any [] media of [their] choice."[3] When a person seeks reliable information, such as accurate medical information during a pandemic[4] or voting information during an election,[5]

---

[2] *Definitions*, THE MEDIA MANIPULATION CASEBOOK, https://mediamanipulation.org/definitions (last visited Feb. 6, 2021).
[3] G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights, Art. 19 (Dec. 16, 1966).
[4] *See* Johnathan Corpus Ong, *Southeast Asia's Disinformation Crisis: Where the State is the Biggest Bad Actor and Regulation is a Bad Word* (Jan. 12, 2021), https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/southeast-asias-disinformation-crisis-where-the-state-is-the-biggest-bad-actor-and-regulation-is-a-bad-word/.
[5] Pam Fessler, *Robocalls, Rumors, and Emails: Last-Minute Election Disinformation Flood Voters*, NPR (Oct. 24, 2020), https://www.npr.org/2020/10/24/927300432/robocalls-rumors-and-emails-last-minute-election-disinformation-floods-voters.

disinformation violates their rights by polluting the information ecosystem with false or misleading ideas that make it harder to access timely, relevant, and accurate information. It is commonly argued that mitigating the spread of such false information presents a challenge to freedom of expression and could lead to censorship. Yet disinformation, if left unchecked, may also become a threat to expression and access to information. Interventions to mitigate the spread of misinformation are therefore needed to protect these basic human rights.

Firstly, disinformation at scale can obscure accurate information, which then undermines the ability to receive accurate information. Secondly, false claims that are amplified widely and quickly through the internet and larger media sphere, can imperil the freedom of expression of those individuals and groups targeted, by silencing them, harassing them, and burying their contributions to the information ecosystem under a miasma of misinformation.[6],[7] Mitigating the spread of disinformation, therefore, is not necessarily at odds with freedom of expression, but may – if done with transparency, oversight, community input and expertise, and the goal to protect and encourage civic participation – actually promote it. If left unchecked, however, it would mean continuing the status quo, which not only prioritizes sensationalism and traffic for profit but the amplification of those who have the most money and resources. The result, in some cases, is real world harm, as seen in the persecution against the Rohingya people in Myanmar,[8] the Capitol Hill siege on January 6, 2021,[9] and the rise of Islamophobia in India.[10]

---

[6] *See* Gina Masullo Chen et al., *'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists*, 21 JOURNALISM 877 (Apr. 2018).
[7] *Toxic Twitter – The Silencing Effect*, AMNESTY INTERNATIONAL, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-5/ (last visited Feb. 15, 2021).
[8] Paul Mozur, *A Genocide Incited on Facebook, With Posts From Myanmar's Military*, NY TIMES (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.
[9] Joan Donovan & Gabrielle Lim, *The Internet Is a Crime Scene* (Jan. 20, 2021), https://foreignpolicy.com/2021/01/20/internet-crime-scene-capitol-riot-data-information-governance/.
[10] Alexandre Capron, *'CoronaJihad': Fake news in India accuses Muslims of deliberately spreading Covid-19* (May 13, 2020), https://observers.france24.com/en/20200513-india-coronajihad-fake-news-muslims-spreading-covid-19.

1.1. The Findability Problem: Misinformation is easily produced and when amplified, can pollute the online ecosystem and impede the right to receive accurate information.

The right to information is a fundamental element of international and regional human rights law.[11] Central to the spirit of this right is the internationally recognized right to the truth,[12] which we interpret as the need for "timely, local, relevant, and accurate" information. However, as news consumers are increasingly moving online, new opportunities for exploitation are available for motivated actors who are intent on seeding false and misleading content. Two key contributing technological factors are (1) the ease with which the internet allows people to share false or misleading information, and (2) the algorithmic amplification of this disinformation.

First, accurate information is often displaced by disinformation during moments of heightened attention to a particular topic. Disinformers and media manipulators often leverage breaking news to set media agendas and generally sow chaos through misidentification or falsification of information. Particularly, on Twitter, disinformation flourishes alongside trending topics, which makes it difficult to find accurate information. On Facebook, disinformation is often seeded into groups and then is shared out on individual pages. As the disinformation campaign reaches more people, it is ranked higher in search and recommendation algorithms, thus reinforcing its prevalence.

Second, recommendation systems and trending algorithms can be gamed to amplify false or misleading content. Media manipulators will create multiple versions of blogs, posts, videos, and images to make the disinformation appear more popular than it really is. Thus, if an ad is repeatedly shown to a user, that user is more likely to take it as the truth.[13] Misinformation peddlers can take advantage of this tendency by creating false or automated accounts to engineer engagement. One reason disinformation is so pervasive is that the tendency to believe information stems not from the content itself, but the source,[14] making deliberately false information "believable" if it is coming from a trusted

---

[11] *See supra* note 3.
[12] G.A. Res. 68/165, Right to the Truth (Dec. 18, 2013).
[13] Emily Dreyfuss, *Want to Make a Lie Seem True? Say It Again. And Again. And Again*, WIRED (Feb. 11, 2017), https://www.wired.com/2017/02/dont-believe-lies-just-people-repeat/.
[14] *Id.*

source.[15] This is particularly problematic online because so much content lacks proper context and provenance.

While there are many tactics available to disinformers, the truth tends to be static and relatively boring. Because social media and search engines optimize based on a set of signals from users and the content itself, an entire industry has flourished around search engine optimization. The SEO industry has pioneered a number of strategies that advantage disinformers over truthful information. One common tactic for spreading misinformation is inauthentically generated support through bots, which are artificial accounts that use anonymized techniques to amplify content.[16,17] Manipulating search through keyword squatting, i.e., mislabeling or miscategorizing disinformation on purpose, has been an especially effective strategy for tethering disinformation to the unique names, locations, or breaking news events. When paired together, the manipulation of engagement and search returns displace the truth.

Algorithmic recommendation systems are particularly vulnerable to these tactics because algorithms neither fact-check information,[18] nor have ethics training like that of a professional journalist or librarian,[19] and thus mix disinformation with accurate information. With every share and retweet this information gets amplified quickly and widely regardless of whether it was someone's intention to spread misinformation. For example, during the U.S. capitol riots the "Capitol Meemaw" meme went viral even though the subject of the meme was not actually present at the riots. [20] This goes to show that wherever there is an opportunity for misinformation to spread, regardless of intention, it will spread and will displace accurate information. Thus, with a system

---

[15] Adam M. Enders et al., *The Different Forms of COVID-19 Misinformation and Their Consequences*, HARVARD KENNEDY SCHOOL MISINFORMATION REVIEW (Nov. 16, 2020), https://misinforeview.hks.harvard.edu/article/the-different-forms-of-covid-19-misinformation-and-their-consequences/.

[16] *See supra* note 2.

[17] Brian Friedberg & Joan Donovan, *On the Internet, Nobody Knows You're a Bot: Pseudoanonymous Influence Operations and Networked Social Movements*, 6 JODS (Aug. 7, 2019).

[18] *Id.*

[19] Joan Donovan & danah boyd, *Stop the Presses? Moving From Strategic Silence to Strategic Amplification in a Networked Media System*, AM. BEHAVIORAL SCIENTIST (2019).

[20] *See* David Mack, We Tracked Down "Capitol Meemaw" — Who Was Not Actually At The US Capitol, BUZZFEED NEWS REPORT (Jan. 12, 2021), https://www.buzzfeednews.com/article/davidmack/capitol-meemaw-meme-topeka-kansas.

tailored towards virality, content curation is essential to stop the spread of misinformation.

1.2. The Silencing Problem: The current online ecosystem enables targeted harassment, which has a chilling effect on freedom of expression.

Algorithms amplify more than just inaccurate information. When the disinformation in question is targeted harassment, amplification has an additional chilling effect on the right to freedom of expression. Although social media companies often have policies that forbid harassment, they are not evenly enforced, and companies may be slow to remove harassment even when it is coordinated as part of a disinformation campaign.[21] As a result, targets of such campaigns turn to self-censorship, either by shutting down their accounts completely, altering the content of their expression, or shifting to less public platforms for communication.[22],[23],[24]

One example of targeted harassment is that used by political partisans to mobilize their supporters and leverage media manipulation techniques for oppressive means.[25] These actors may use short and catchy phrases, known as viral slogans,[26] coupled with a coordinated effort to spread the message, a tactic known as swarming,[27] to silence dissent.[28] Often this takes the form of thousands of people posting the same hateful comment on their target's social media accounts. This technique has been used to swarm women online in underserved Nigerian communities and by "virtual lynch mobs" in Turkey to push its targets into self-censorship.[29] Actors use these techniques to flood their targets with violent comments online if they speak out online.[30] Absent mitigation,

---

[21] Chen, *supra* note 6.
[22] Andreas Reventlow, *The chilling effects of online harassment and how to respond* (Dec. 6, 2016), https://www.mediasupport.org/chilling-effects-online-harassment-address/
[23] GABRIELLE LIM, SECURITIZE/COUNTER-SECURITIZE THE LIFE AND DEATH OF MALAYSIA'S ANTI-FAKE NEWS ACT, https://datasociety.net/wp-content/uploads/2020/04/Securitize-Counter-securitize.pdf.
[24] *Troll Patrol Findings*, AMNESTY INTERNATIONAL, https://decoders.amnesty.org/projects/troll-patrol/findings#what_did_we_find_container (last visited Feb. 15, 2021).
[25] *See* Anthony Nadler et al., *Weaponizing the Digital Influence Machinery The Political Perils of Online Ad Tech* (Oct. 17, 2018), https://datasociety.net/library/weaponizing-the-digital-influence-machine/.
[26] *See supra* note 2.
[27] *Id*.
[28] *See* Ong, *supra* note 4.
[29] *See* Reventlow, *supra* note 22.
[30] Chen, *supra* note 6.

this form of amplified disinformation can have a direct chilling effect on freedom of expression.[31]

Compounding the harm, targeted harassment online is frequently directed towards people who have already been marginalized,[32] arguably the very people whom human rights law was created to protect. This is happening around the world, where targeted harassment has been used against vulnerable groups to crack down on political dissent,[33] and in some cases are largely led by state actors themselves against their minority populations or political opponents.[34] In the Philippines, for example, the arrest of journalist Maria Ressa follows years of intimidation and harassment by supporters of the ruling party.[35],[36] While she has remained an outspoken figure despite the threats, this climate of fear and intimidation has resulted in a culture of burnout, fear, and self-censorship among the wider media industry in the Philippines.[37],[38] Thus, if diverse and inclusive opinions online are a concern,[39] then allowing amplified disinformation to go unmitigated has a direct silencing effect on the opinions that the UN seeks to protect.

---

[31] JON PENNEY, *Online Abuse, Chilling Effects, and Human Rights*, *in* CONNECTED CANADA: A RESEARCH AND POLICY AGENDA (E. Dubois & F. Martin-Bariteau eds., 2020).
[32] *Id*.
[33] *See* Soma Basu*, Manufacturing Islamophobia on WhatsApp in India,* THE DIPLOMAT (May 10, 2019), https://thediplomat.com/2019/05/manufacturing-islamophobia-on-whatsapp-in-india/.
[34] *See* Ronan Lee, *Extreme Speech| Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis*, 13 INT'L J. COMMC'N (2019).
[35] Hannah Ellis-Petersen, *Maria Ressa: editor of Rappler news website arrested on 'cyber-libel' charges*, THE GUARDIAN (Feb. 13, 2019), https://www.theguardian.com/world/2019/feb/13/philippines-journalists-decry-intimidation-as-website-editor-arrested.
[36] Heather Timmons, *Maria Ressa's arrest is a warning to every journalist in a democracy*, QUARTZ (Feb. 13, 2019), https://qz.com/1549538/maria-ressas-arrest-by-rodrigo-duterte-is-a-warning-to-every-journalist/.
[37] Sheila S. Coronel, *A 'Fraught Time' For Press Freedom in The Philippines,* NPR (Jan. 17, 2018), https://www.npr.org/sections/parallels/2018/01/17/578610243/a-fraught-time-for-press-freedom-in-the-philippines.
[38] *CPJ mission finds increased intimidation, shrinking space for free press in the Philippines*, COMMITTEE TO PROTECT JOURNALISTS (Apr. 15, 2019), https://cpj.org/2019/04/cpj-mission-finds-increased-intimidation-shrinking/.
[39] Reventlow, *supra* note 22.

2. HOW TO AMPLIFY THE TRUTH: MOVING TOWARDS A PUBLIC INTEREST INTERNET

The debate over how to stop disinformation includes several proposed remedies to ensure compliance to site policies, local norms, and the law.[40] In this debate, two words often come to mind: moderation and curation. Moderation is often what people call for when asking governments and companies to remove harmful and false posts from the public sphere. But over a decade of social media has revealed that this approach is not sufficient.[41] *Instead, we recommend working toward community-based content curation, implemented as a proactive course of action that can be complemented by moderation.*

Content moderation is the reactive process of reviewing and deciding whether content created by a user is objectionable to the online community or in violation of a specific website's terms of service.[42] Its origins go back to the online forums of the 1970s, when most moderation was done by volunteers to ensure the discussions followed certain rules and to prevent inappropriate topics, discussions, and content from being shared within the community.[43] Issues such as the liability of content moderators emerged during the 1990s,[44] and many companies have delegated the responsibilities regarding content moderation to third parties. Currently in many countries, underpaid workers are tasked with viewing this harmful content and deciding what to delete.[45] In addition to being emotionally and mentally taxing on the moderators, it also doesn't work consistently.

Firstly, moderation is limited to reacting post hoc to content that has already been shared. This renders it ineffective at preventing many instances of mis- and disinformation from being seen and shared widely as moderation is often slow, inconsistent, incremental, or merely ineffective.[46] Many technology companies in the United States, for example, recently added new content moderation policies,[47] such as labeling misleading content,

---

[40] CHUNG SHENG-LI ET AL., NEW FRONTIERS IN COGNITIVE CONTENT CURATION AND MODERATION 3, (Cambridge University Press ed. July 23, 2018), available at https://www.cambridge.org/core/journals/apsipa-transactions-on-signal-and-information-processing/article/new-frontiers-in-cognitive-content-curation-and-moderation/DF4AAE1F2052DF784E52B7882208AF15.
[41] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (Yale University Press ed. 2019).
[42] *Id.* at 3.
[43] *Id.* at 2.
[44] *Id*.
[45] *Id*.
[46] *Id.*
[47] *Id*.

publishing transparency reports of "coordinated inauthentic behavior,"[48] or redirecting users to more credible and authoritative content.[49] However, like fact-checking and media literacy, the effectiveness of these measures is still up for debate and often happen long after the content has already been widely shared. Labeling misleading or false content, for example, may backfire as it may imply that anything without a label is true,[50] while banning users or removing content may simply shift those users and the content to other platforms.[51] Furthermore, social media platforms are increasingly outsourcing content moderation to companies that are ill-equipped to understand regional contexts,[52] but have the effect of releasing the company from liability for harassment, incitement, and hate.[53]

In addition, moderation solutions implemented by technology companies are often at risk for increased censorship and surveillance. In China, for example, content policies are typically handed down by the Chinese Communist Party (CCP), but it is the private companies who are responsible for carrying out the content moderation. The result is undue censorship, as companies are incentivized to over-correct, so they do not violate the CCP's directives.[54] And in countries where there is local legislation that criminalizes false information, content removal and arrests have been common. In Singapore, for example, the Protection from Online Falsehoods and Manipulation Act, has resulted in

---

[48] Amelia Acker & Joan Donovan, *Data Craft: A Theory/Methods Package for Critical Internet Studies*, 22 INFO., COMMC'N & SOCIETY 1590 (2019).
[49] Clea Skopeliti & Bethan John, *Coronavirus: How Are the Social Media Platforms Responding to the 'Infodemic'?,* FIRST DRAFT (Mar. 19, 2020), https://firstdraftnews.org:443/latest/how-social-media-platforms-are-responding-to-the-coronavirus-infodemic/.
[50] Gordon Pennycook et al., *The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings*, 66 MGMT. SCI. 4921 (Feb. 2020).
[51] P. M. Krafft & Joan Donovan, *Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign*, 37 POLITICAL COMMC'N 194 (2020).
[52] MARGARET E. ROBERTS, CENSORED (Princeton University Press ed. 2020), https://press.princeton.edu/books/hardcover/9780691178868/censored.
[53] JOAN DONOVAN & GABRIELLE LIM, DETECT, DOCUMENT, AND DEBUNK: STUDYING MEDIA MANIPULATION AND DISINFORMATION (Oxford Handbook).
[54] Lotus Ruan et al., *The Intermingling of State and Private Companies: Analysing Censorship of the 19th National Communist Party Congress on WeChat*, THE CHINA QUARTERLY (July 2020).

Facebook labeling content the government deems to be false – an act that has been widely criticized by human rights groups and opposition politicians[55,56]

Moreover, illiberal and authoritarian-leaning governments have used disinformation as a pretense to crack down on dissent.[57] In Egypt, for example, the government justifies arresting and intimidating regime critics and other forms of digital expression as safeguarding national security from "false information."[58] Even within established democracies, the fear of "foreign speech" has similarly raised concerns over potential infringements on freedom of expression and the further balkanization of the internet.[59]

Community-based content curation, on the other hand, is proactive. Curation is the act of collecting, sorting, and organizing community-generated content around a topic and actively promoting the most useful, timely, and accurate information. Content curation was first proposed as a solution to the challenge of organizing online content during the early 1990s, by the Digital Library Initiative[60] and again in the early 2000s by Tim Berners-Lee's semantic web.[61,62]

Curators choose to highlight the best content based on quality, not on popularity. Librarians, for example, are professionally trained to identify trustworthy sources and contents that should be available online.[63] In addition to ensuring compliance within a certain framework, curation also aims for accuracy and relevance. *While a moderator checks whether content is acceptable under a set of rules, a curator selects the most useful, timely, and accurate content in order to display relevant information for users. We*

---

[55] Rachel Au-Yong, *Parliament: Workers' Party Opposes Proposed Law on Fake News, Says Pritam Singh*, THE STRAITS TIMES (May 7, 2019), https://www.straitstimes.com/politics/parliament-workers-party-opposes-proposed-law-on-fake-news-pritam-singh.
[56] *RSF Explains Why Singapore's Anti-Fake News Bill Is Terrible*, REPORTERS WITHOUT BORDERS (Apr. 8, 2019), https://rsf.org/en/news/rsf-explains-why-singapores-anti-fake-news-bill-terrible.
[57] Elana Beiser, *Hundreds of Journalists Jailed Globally Becomes the New Normal, Committee to Protect Journalists* (Dec. 13, 2018), https://cpj.org/reports/2018/12/journalists-jailed-imprisoned-turkey-china-egypt-saudi-arabia/.
[58] *Id*.
[59] Gabrielle Lim, *The Risks of Exaggerating Foreign Influence Operations and Disinformation*, CENTRE FOR INT'L GOVERNANCE INNOVATION (Aug. 7, 2020), https://www.cigionline.org/articles/risks-exaggerating-foreign-influence-operations-and-disinformation.
[60] Edward A. Fox & Ohm Sornil, *Digital libraries*, ENCYCLOPEDIA OF COMPUTER SCIENCE 576 (2003).
[61] Tim Berners-Lee et al., *The Semantic Web*, 284 SCI. AM. 28 (2001).
[62] Sheng-Li, *supra* note 40.
[63] *See* Joan Donovan, *You Purged Racists From Your Website? Great, Now Get to Work*. (July 1, 2020), WIRED, https://www.wired.com/story/you-purged-racists-from-your-website-great-now-get-to-work/.

recommend complementing content moderation with active curation. Where curation fails, moderation can step in, but both methods must work together.

Curation has been growing in several practical perspectives. For example, in terms of business models, Facebook has implemented an initiative called News Tabs, a new section inside of the company's mobile application that will surface the most recent and relevant stories for readers.[64] Instead of relying on algorithms to filter information, the company hired journalists and reporters to filter the best content. In general, social media companies might step up to the challenge and build a content curation model for search, trends, and recommendation that does not rely so heavily on reactionary moderation.[65]

To avoid some of the problems attendant to moderation, such as influence from political elites and censorship, we stress that curation should be *community-focused and grounded in community input and expertise and with the goal to protect and encourage civic participation*. This means having humans in the loop and not delegating all content curation to algorithmic systems. Furthermore, it requires community input and a bottom-up approach that puts safety, trust, and transparency at the forefront — not traffic or profit.

3. RECOMMENDATIONS

Recognizing the particular challenges posed by mis- and disinformation, the role of algorithmic content curation and propagation, and the potential for *community-based content curation* to address those challenges is an important first step in protecting freedom of expression. Once that step is taken, it will be incumbent on governments, content platforms, media outlets, and other stakeholders to follow through with concrete action. While a full analysis of the path from moderating misinformation to curating information is beyond the scope of this comment, there are several promising steps that would put us closer to a functioning community-based curation strategy. We must promote:

---

[64] Mike Isaac, *In New Facebook Effort, Humans Will Help Create Your News* Stories, N.Y. TIMES (Aug. 20, 2019), https://www.nytimes.com/2019/08/20/technology/facebook-news-humans.html.
[65] Joan Donovan, *Combatting the Cacophony with Librarians,* GLOBAL INSIGHTS (Jan. 2021), https://www.ned.org/wp-content/uploads/2021/01/Combating-Cacophony-Content-Librarians-Donovan.pdf.

- *Transparency*: In order to implement effective content curation, it is important to first know who is currently in control of content and what practices they currently use to shape it. This transparency will not only help the public understand who to hold accountable, but also will help to identify where in the process improvements can be made. Governments should publish policies relevant to content regulation online, identifying any orders issued to social media platforms and other online content providers.[66] At a bare minimum, content providers should make their content restriction policies, decision making processes, and actions available online and in plain language.[67] Ideally, they should go a step further in creating transparency by publishing the algorithms they use for content moderation online, as app developers did when developing contact tracing apps during the COVID-19 pandemic.[68] Open access to algorithms' source fosters improvement through a participatory public process,[69] and additionally facilitates the replication of successful algorithms.[70] Lastly, all advertising should be clearly labelled and traceable back to the purchaser.

- *Durability*: Search and content recommendation algorithms allow content providers to react to individual users' activity – but this reactivity can be easily gamed. Search engines and social media platforms should build durability into their functionality to prevent keyword squatting and other attacks.[71] This means making platforms less reactive to small changes and more responsive to long-term advantages of a stable information ecosystem, including trends that are identified by trained curators rather than algorithms.

- *Building in multi-stakeholder engagement into development*: Content moderation – the reactive removal of harmful content – arguably stretches the expertise of platform developers. *Community-based content curation* – proactive promotion of useful, contextual, truthful information – is likely beyond that expertise entirely.

---

[66] *Manila Principles on Intermediary Liability*, https://www.manilaprinciples.org/ (last visited Feb. 8, 2021).
[67] *Id*.
[68] *Open Source Solutions*, DIGITAL RESPONSE TO COVID-19, https://joinup.ec.europa.eu/collection/digital-response-covid-19/open-source-solutions (last visited Feb. 8, 2021).
[69] *The Power of Open Source AI*, FORBES (May 22, 2019), https://www.forbes.com/sites/insights-intelai/2019/05/22/the-power-of-open-source-ai/?sh=4b1031276300.
[70] *See* Tom Bianchi, *Open Source Should Always Have Been the Way for COVID-19 Contact Tracing Apps*, CITY A. M. (Sept. 17, 2020), https://www.cityam.com/open-source-should-always-have-been-the-way-for-the-covid-19-contact-tracing-app/.
[71] *See supra* note 2.

However, there are professionals, such as librarians, civil society organizations, journalists, and other stakeholders who have that expertise.[72] Technology companies therefore need to understand the limits of their own capabilities and build in multi-stakeholder engagement into their development process and throughout the life of their product – in other words, hiring and working with individuals who are trained and qualified to curate content and systematically privilege credible and responsible voices over inflammatory, divisive, sensational content. Reddit is organized by communities known as subreddits and who foster a bottom-up approach to curation that is driven not by algorithms but by the members of each community.[73] However, all subreddits are subject to the rules of the platform. Alternatively, social media companies might step up to the challenge by hiring librarians to build a content curation model that does not rely so heavily on reactionary moderation.[74]

- *Creating infrastructure that encourages democratic participation and accountability* – Curation and moderation policies and enforcement, while important, are not enough. How platforms organize information and groups also impacts whether mis- or disinformation is readily spread. New information technologies should therefore also consider how best to build network infrastructure that allows individuals and communities to engage in ways that promote democratic participation and prioritizes authenticity, legibility, and accuracy. While this area of research is still nascent, we encourage further research drawing from archival studies, infrastructure studies, library science, network science, and organizational sociology.[75],[76],[77]

While there is no communication without the presence of some misinformation, this should not be the guiding principle for our global information commons. In fact, just as media manipulators depend on journalists to cover both sides of a story and took advantage of that ethic to garner unearned attention, disinformers depend on the

---

[72] Donovan, *supra* note 63.
[73] Jennifer Forestal, *Beyond Gatekeeping: Propaganda, Democracy, and the Organization of Digital Publics*, 83 J. POL. (Jan. 2021).
[74] Donovan, *supra* note 65.
[75] Forestal, *supra* note 73.
[76] Joan Donovan, *Navigating the Tech Stack: When, Where and How Should We Moderate Content?*, CENTRE INT'L GOVERNANCE INNOVATION (Oct. 28, 2019), https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content.
[77] Susan Leigh Star, *The Ethnography of Infrastructure*, 43 AM. BEHAV. SCI. (Nov. 1999).

inaction of technology companies to ensure their campaigns go viral. Throughout the last five years, researchers have documented the same pattern: well-funded groups with large follower networks across platforms leveraged breaking news to control media agendas, especially on topics related to race, public health, politics, and gender. Much can be done to prevent misinformation from reaching millions. Like secondhand smoke, misinformation-at-scale damages the quality of public life and over time has a corrosive effect on our society. Therefore, we must make it more difficult to spread mis- and disinformation-at-scale and offer up a new vision for a public interest internet that takes community safety as its most valued feature.