PART 1

DIGITAL PLATFORM GOVERNANCE AND THE CHALLENGES FOR TRUST AND SAFETY

RESEARCH ICT AFRICA

WHY LIES AND HATRED PROLIFERATE ON DIGITAL PLATFORMS

KEY TRENDS UNCOVERED

- Online and platform content that may cause harm through the breach of human rights is sufficiently widespread to have raised concerns about the potentially severe implications for the future of trust, safety, democracy and sustainable development
- A certain amount of this content is curbed by the dominant commercial platforms' content moderation mechanisms. Much still escapes their nets and in worst cases is algorithmically amplified and even supported by advertising.
- Some smaller platforms expressly allow hatred and conspiracy theories, even facilitating the organisation of offline attacks on democracy.
- The roots of the problems lie in : 'attention economics', automated advertising systems, external manipulators, company spending priorities and stakeholder knowledge deficits.
- Of value in addressing these problems will be the development of guidelines for regulating platforms, centred on safeguarding human rights, promoting transparency and limiting the business processes and technical mechanisms that underpin potentially harmful content online.

*This is a draft background paper developed with the support of UNESCO by Research ICT Africa, a digital policy, regulation and g*overnance think tank, based in Cape Town, South Africa. The designations employed and the presentation of material throughout do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The ideas and opinions expressed are those of the authors; they are not necessarily those of UNESCO and do not commit the Organisation.

This is Part 1 of a three-part series. The consolidated report is an evidence-based input to the consultative processes in UNESCO's project titled "Guidance for regulating digital platforms: a multistakeholder approach". Each Part stands alone but can also be profitably read as an element in the series.

- Part 1 tackles the what and the why about problems in platform content
- Part 2 deals with the how, with a focus on platforms' policies and practices
- Part 3 looks at possible solutions through diverse regulatory arrangements

Evidence reviewed in this brief rests on work by the academic, civil society and journalistic communities, as well as on documents from the platforms themselves. More than 800 documents, mainly published between 2020 and 2022, were identified and assessed with a view to current debates about regulatory frameworks.

1 The importance of a healthy information environment for democracy

This is Part 1 of a draft background paper that analyses how the current state of platform governance accounts for harms to human rights arising from online content. Acknowledging that the term "platform" is not a neutral one[1], and that there is no universally accepted definition, it should be noted that the focus below is on the content. A deeper assessment of the entities that transmit this content, the platforms, is in the third Part in this series.

Amongst other diagnoses of harms linked to digital platforms are key UN resolutions[2] which have underlined negative impacts on safety and trust through the spread of online disinformation, misinformation and hate speech. The resolutions have noted how this content can undermine human rights and sustainable development. To this can be added the recognition that problematic expression is often deployed to harass marginalised groups[3], and in the process also fuses a range of intersectional features like skin colour, ethnicity, religion, sexual orientation, and nationality[4].

As an indication of growing concerns with online content, UNESCO-commissioned research by the Social Media Insights Lab at the University of Carolina in 2022 found a significantly increasing trend in online mentions of the terms "fake news", "misinformation" and "disinformation" (the two latter terms especially since 2020). This trajectory emerges from an analysis of a dataset of 280 million references since 2012, covering Facebook, YouTube, Twitter and Google search.

Illustrating that online content governance is an issue beyond the prominent Western platforms, a global inventory of the most-used 100 social media (in terms of monthly active users) has identified that although US-governed platforms are dominant, 61% of the other companies in the list include large Chinese players (within the top ten), as well as other enterprises with high significance at the regional or country-level.[5]

Research shows that many of these platforms carry what has been dubbed a "payload" of "adversarial narratives".[6] Evidence also reveals that content which can enable harm to human rights is present even on small, and sometimes overlooked, communications platforms. For example, a dataset of 183 million Parler posts from four million users shows that this particular platform allowed conspiracy narratives and violent extremist groups, and facilitated coordination for storming of the U.S. capitol on January 6, 2021.[7]

---

1 Gillespie, Tarleton, 'The Politics of 'Platforms'', New Media & Society, Vol. 12, No. 3, 2010, Available at SSRN: https://ssrn.com/abstract=1601487

2 UN General Assembly, 'Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms', 2021

3 Caroline Sinders, 'The Use of Mis- and Disinformation in Online Harassment Campaigns', *Center for Democracy and Technology* (blog), 2022, https://cdt.org/insights/the-use-of-mis-and-disinformation-in-online-harassment-campaigns/.

4 Md Sayeed Al-Zaman, 'Digital Disinformation and Communalism in Bangladesh' (SocArXiv, 2019), https://doi.org/10.31235/osf.io/8s6jd. flourishing consumer culture, and security in social life as a cumulative force smooths the scope of modern global amenities to come in and grow up amid this changing society. Of them, new age digital communication is vital one. Digital media is encompassing people's everyday life. Process of acquiring information has also changed remarkably: instead of searching to get one, people now struggle to look for reliable information due to ample information. Cyberspace becomes the cornucopia of fluid information that often baffles the surfers by providing distorted information. Bangladesh has been experiencing digital media-initiated disinformation from the beginning of 2010s. Interest groups are playing with digital disinformation conjoining religious sentiment. As a result, incidents of assault on religious minorities based on digital (dis

5 Chand Rajendra-Nicolucci and Ethan Zuckerman, *An Illustrated Field Guide to Social Media*, 2021, 2021, http://knightcolumbia.org/blog/an-illustrated-field-guide-to-social-media; Wikipedia, 'List of Social Platforms with at Least 100 Million Active Users', in *Wikipedia*, 10 December 2022, https://en.wikipedia.org/w/index.php?title=List_of_social_platforms_with_at_least_100_million_active_users&oldid=1126634532.

6 GDI, 'The Global Disinformation Index', 2019, https://www.disinformationindex.org/.

7 Max Aliapoulios et al., 'An Early Look at the Parler Online Social Network' (arXiv, 18 February 2021), http://arxiv.org/abs/2101.03820.

It is also evident that proponents of potentially harmful content ply their trade across several platforms (and languages) to secure the widest reach possible.[8] Meanwhile, most people access on average seven platforms a month[9], and so content issues are overlapping[10]. Further, it is evident that actors interested in causing harm to human rights continue this role on the smaller platforms when experiencing restrictions or 'deplatforming' on the larger ones.[11] Closed messaging systems and groups can also have serious consequences for human rights, and they often cascade problematic content into more public fora.[12] While bigger platforms may logically carry more responsibility than smaller ones, it is important to take a comprehensive perspective of the wider platform landscape.

Content harms linked to platforms can be significant in those services whose principal purposes diverge from the classic functions of social media postings. For instance, on the international gaming platform Twitch, a search uncovered 73 videos and 91 channels that expressed support for extreme right-wing ideologies, and which risked livestreaming of human rights abuse.[13] Another platform, DLive, dedicated to streaming, was found to have at least 100 extreme right accounts used by influencers to reach established audiences and strengthen existing extremist communities.[14]

All of this is context for understanding the proliferation of online narratives that constitute (and often combine)[15] online hate speech, misinformation and disinformation – with major significance for democracies, and particularly electoral systems, as well as for public health and safety, sustainable development and the prospects for advancing UNESCO's interests in "information as a public good[16].

ON THE GROUND: In conflict conditions, the stakes of what occurs online in relation to disinformation are enormous. A court case in late 2022 related to the November 2021 assassination in Ethiopia of Prof Meareg Amare Abrha as he was trying to enter the family home. According to his son, Facebook posts before the attack had slandered him and revealed identifying information. The BBC reported that despite repeated complaints using Facebook's reporting tool, the posts had stayed up. The professor's son accused the platform of being "woefully inadequate", with too few moderators who deal with posts in key languages Amharic, Oromo and Tigrinya.[17]

8 Moustafa Ayad, Anisa Harrasy, and Mohammed Abdulla, A, 'Under-Moderated, Unhinged and Ubiquitous: Al-Shabaab and the Islamic State Networks on Facebook', 2022.

9 Kemp, 'Digital 2022'.

10 Tarleton Gillespie and Patricia Aufderheide, 'Expanding the Debate About Content Moderation', SSRN Scholarly Paper (Rochester, NY, 19 May 2020), https://papers.ssrn.com/abstract=3629185.

11 Richard Morris, 'Researchers warn of rise in extremism online after Covid', 2022, https://www.bbc.com/news/uk-politics-61106191

12 Jacob Gursky et al., 'Chat Apps and Cascade Logic: A Multi-Platform Perspective on India, Mexico, and the United States', *Social Media + Society* 8, no. 2 (1 April 2022): 20563051221094772, https://doi.org/10.1177/20563051221094773.

13 O'Connor, 'Digital Activism and the Increased Role of Dalit Activism in Intersectional Feminism in India'; UNOCT, 'Examining the Intersection Between Gaming and Violent Extremism', 2022.I argue that the rise of social media has given Dalit activism unprecedented levels of visibility and brought Dalit emancipatory politics into the repertoire of intersectional feminism in India. This article will examine the ways in which social media has changed social movements and given voice to Dalit activism in India through analyzing anti-harassment feminist hashtag activism, particularly the #WhyLoiter movement, representation of Dalit women in mainstream media and subsequent social media response, and, finally, the ways in which the former two intersect with Dalit emancipatory politics in India. The concluding sentiments of this article are that social media has given way to subaltern views of social justice, in which the needs of women and Dalit communities (and Dalit women! https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/221005_research_launch_on_gaming_ve.pdf

14 Elise Thomas, 'The Extreme Right on DLive', https://www.isdglobal.org/wp-content/uploads/2021/08/03-gaming-report-dlive-1.pdf

15 Jae Yeon Kim and Aniket Kesari, 'Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic | Journal of Online Trust and Safety', 28 October 2021, https://tsjournal.org/index.php/jots/article/view/13.

16 UNESCO, 'Windhoek+30 Declaration. Information as a public good', 2021, https://en.unesco.org/sites/default/files/windhoek30declaration_wpfd_2021.pdf

17 Killing in Ethiopia', BBC News, 14 December 2022, sec. Technology, https://www.bbc.com/news/technology-63938628; See also Peter Mwai, 'Ethiopia's Tigray Conflict: What Are Facebook and Twitter Doing about Hate Speech? - BBC News', accessed 16 December 2022, https://www.bbc.com/news/59251942.

While figures on the reach of posts that may violate human rights do not equate to actual impact, the potential problems can be exacerbated by scale as several recent findings illustrate:

- Misogyny[18] and racist[19] expression have found widespread traction on the platforms: in one given period, researchers identified almost seven million instances of online hateful speech against women, LGBTQ communities, people with disabilities and French Arab communities.[20]
- As revealed by UNESCO research, a sample of content about the Holocaust on Telegram public groups revealed that almost half the selection consisted of denial and distortion. Equivalent content still constituted a fifth of samples studied on Twitter and on TikTok.[21] Comments on TikTok against Jewish content creators have been left online on the same platform. Six fringe platforms failed altogether to address antisemitic content.[22]
- In five months in 2022, South Africa saw more than three million mentions of xenophobic content on social media, up 130% from the previous period.[23]
- In the European elections of 2019, 500 pages and groups on Facebook promoting disinformation received 533 million views and were liked, commented upon or shared by 67 million people.[24]
- In a study of the 2019 Indian general election, researchers joined 600 WhatsApp groups and found evidence of 75 manipulation campaigns in the form of mobilisation messages with lists of pre-written tweets.[25]
- Research from the USA's 2020 elections shows that more than 300 distinct 'stories' or narratives, encompassing 44.8 million tweets, sought to sow doubt about the poll.[26]

Such problematic content seeks to discredit and displace what is truthful, thereby polluting the public's right to know. It is often used to incite hatred and violence, and to intimidate others into self-censoring.[27] While some companies employ 'friction' (such as limiting engagement, redirecting search results, and/or adding context or labels), the general picture is one where "repeat spreaders" of problematic expression, even when due process (like three warnings) is considered, are seldom sanctioned. The overall result is that the flow of misleading and hateful online narratives simply continues.[28]

Among other potential harms, the corruption of information – in text, sound and image - correlates with the erosion of credibility and integrity of elections in many countries and therefore is likely to undermine the right to political participation.

18 Elise Thomas, 'The Extreme Right on DLive', 2021; Ellen Judson, 'Gendered Disinformation: 6 Reasons Why Liberal Democracies Need to Respond to This Threat', Demos, accessed 18 December 2022, https://demos.co.uk/blog/gendered-disinformation/; Ellen Judson et al., ' Engendering Hate: The contours of state-aligned gendered disinformation online, https://demos.co.uk/wp-content/uploads/2022/02/Engendering-Hate-Oct.pdf

19 'Home Page | Kick It Out', accessed 18 December 2022, https://www.kickitout.org/; Sharon Kimathi, 'Racism in Football: How Will Facebook, Twitter Stop Online Abuse?', news.trust.org, 2021, https://news.trust.org/item/20210712171437-krcu7/.

20 Cooper Gatewood et al., 'Cartographie de La Haine En Ligne', 2019, https://www.isdglobal.org/wp-content/uploads/2019/12/Cartographie-de-la-Haine-en-Ligne-eng.pdf.

21 United Nations and UNESCO, 'History under Attack: Holocaust Denial and Distortion on Social Media', 2022, https://unesdoc.unesco.org/ark:/48223/pf0000382159/PDF/382159eng.pdf.multi.

22 Gabrielle Beacken, Inga Trauthig, and Samuel Wolley, 'Platforms' Efforts to Block Antisemitic Content Are Falling Short', Centre for International Governance Innovation, July 2011, https://www.cigionline.org/articles/platforms-efforts-to-block-anti-semitic-content-are-falling-short/.

23 Centre for Analytics and Behavioural Change. 'An overview of conversations about foreign nationals on South African social media'. June 2022, https://cabc.org.za/wp-content/uploads/2022/06/Xenophobia-in-South-Africa-2022-research-report.pdf

24 Dominik Steiger, 'Protecting Democratic Elections Against Online Influence via "Fake News" and Hate Speech – The French Loi Avia and Loi No. 2018-1202, the German Network Enforcement Act and the EU's Digital Services A…' (Theory and Practice of the European Convention on Human Rights, Nomos Verlagsgesellschaft mbH & Co. KG, 2021), 165–214, https://doi.org/10.5771/9783748923503-165; Beata Martin-Rozumiłowicz and Rasťo Kužel, 'Social Media, Disinformation and Electoral Integrity', 2019.

25 Maurice Jakesch et al., 'Trend Alert: A Cross-Platform Organization Manipulated Twitter Trends in the Indian General Election', *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (18 October 2021): 379:1-379:19, https://doi.org/10.1145/3479523.

26 Justin Hendrix, 'Researchers Release Comprehensive Twitter Dataset of False Claims About The 2020 Election', Tech Policy Press, 15 June 2022, https://techpolicy.press/researchers-release-comprehensive-twitter-dataset-of-false-claims-about-the-2020-election/.

27 Robert Faris and Robert Donovan, 'The Future of Platform Power: Quarantining Misinformation', in *Journal of Democracy*, vol. 32, 3 vols, 2021, 152–56, https://www.journalofdemocracy.org/articles/the-future-of-platform-power-quarantining-misinformation/.

28 Election Integrity Partnership, 'The Long Fuse: Misinformation and the 2020 Election', 2021, https://stacks.stanford.edu/file/druid:tr171zs0069/EIP-Final-Report.pdf; Tania King et al., 'Reordering Gender Systems: Can COVID-19 Lead to Improved Gender Equality and Health?', *The Lancet* 396, no. 10244 (2020): 80–81; Laura O'Connor, 'Digital Activism and the Increased Role of Dalit Activism in Intersectional Feminism in India', *Undergraduate Journal of Politics, Policy and Society* 3, no. 1 (2020): 134–55.\\uc0\\u8216{}Reordering Gender Systems: Can COVID-19 Lead to Improved Gender Equality and Health?\\uc0\\u8217{}, {\\i{}The Lancet} 396, no. 10244 (2020

Among other potential harms, the corruption of information – in text, sound and image[29] - correlates with the erosion of credibility and integrity of elections in many countries[30] and therefore is likely to undermine the right to political participation[31].

There is also an economic cost to societies of online lies and hatred. It affects how a society decides to allocate resources and it diverts spending that could otherwise be channelled into the goals of sustainable development. For example, in the UK it has been estimated that annual social cost of online harm (i.e. the combination of direct costs to victims and society, and indirect costs of worsened mental health and lost productivity) may be at least £13 billion per annum. A component of this is how misinformation relating to COVID-19 and mask-wearing could have weakened the UK economy by £3.6 billion during 2020 through increased caseloads and hospitalisations.[32]

Against this backdrop, there is growing impetus to reconsider the governance status quo which has left it largely optional for the platforms to address these problems. There is growing evidence that harms to human rights, both directly online and with offline impact, are associated with gaps in the wider existing national and international regulatory regimes within which these companies operate. In particular, the common situation of very limited legal liability for platforms in many jurisdictions, is increasingly up for debate.

It is these lacunae that can be addressed by means of a guidance system for regulatory frameworks. Such a system would avoid legitimating legal over-restrictions on platforms, and instead work to protect human rights and advance information as a public good. A guidance system in this vein should not be limited to the aim of curbing content that can harm human rights, but also positively foster legitimate expression and information as a public good. Such a framework would further highlight that regulation does not equate to direct and exclusive control by state entities but encompasses an array of different regulatory arrangements. Examples are the spheres of industry self-regulation, co-regulation mechanisms, and direct statutory regulation, with variations in the legal status and remit of each.

There is growing evidence that harms to human rights, both directly online and with offline impact, are associated with gaps in the wider existing national and international regulatory regimes within which these companies operate.

29 Hana Matatov, Mor Naaman, and Ofra Amir, 'Stop the [Image] Steal: The Role and Dynamics of Visual Content in the 2020 U.S. Election Misinformation Campaign', *Proceedings of the ACM on Human-Computer Interaction* 6, no. CSCW2 (7 November 2022): 1–24, https://doi.org/10.1145/3555599; Himanshu Zade et al., 'Auditing Google's Search Headlines as a Potential Gateway to Misleading Content: Evidence from the 2020 US Election', *Journal of Online Trust and Safety* 1, no. 4 (20 September 2022), https://doi.org/10.54501/jots.v1i4.72.

30 Barrett, Hendrix, J, and Sims, 'Fueling the Fire: How Social Media Intensifies U.S. Political Polarization — And What Can Be Done About It'; Avaaz, 'Facebook. From Election to Insurrection. How Facebook Failed Voters and Nearly Set Democracy Aflame.', 2021; Tech Transparency Project, 'Facebook Profits from White Supremacist Groups', Tech Transparency Project, 10 August 2022, https://www.techtransparencyproject.org/articles/facebook-profits-white-supremacist-groups; CNN, 'Twitter Says It Has Quit Taking Action against Lies about the 2020 Election | CNN Politics', 28 January 2022, https://edition.cnn.com/2022/01/28/politics/twitter-lies-2020-election/index.html.

31 Roman Adamczyk, Sasha Morinière, and Cécile Simmons, 'Le spectre de la fraude électorale : l'impact des discours de désinformation pendant les élections de 2022', 2022; Kantar Public, 'Présidentielle 2022 : Doit-on Craindre Un Impact de La Désinformation Sur La Confiance Dans La Sincérité Du Scrutin ?', Kantar Public, Avril 2022, https://kantarpublic.com/fr/inspiration/election-presidentielle-2022/presidentielle-2022-doit-on-craindre-un-impact-de-la-desinformation-sur-la-confiance-dans-la-sincerite-du-scrutin; Krisztina Rozgonyi, 'THE IMPACT OF THE INFORMATION DISORDER (DISINFORMATION) ON ELECTIONS', 2018; Nic Newman, 'Overview and Key Findings of the 2022 Digital News Report', 15 June 2022, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/dnr-executive-summary.2022; Kantar Public, \\uc0\\u8216{}Pr\\uc0\\u233{}sidentielle 2022\\uc0\\u8239{}: Doit-on Craindre Un Impact de La D\\uc0\\u233{}sinformation Sur La Confiance Dans La Sinc\\uc0\\u233{}rit\\uc0\\u233{} Du Scrutin\\uc0\\u8239{}?\\uc0\\u8217{}, Kantar Public, Avril 2022, https://kantarpublic.com/fr/inspiration/election-presidentielle-2022/presidentielle-2022-doit-on-craindre-un-impact-de-la-desinformation-sur-la-confiance-dans-la-sincerite-du-scrutin; Krisztina Rozgonyi, \\uc0\\u8216{}THE IMPACT OF THE INFORMATION DISORDER (DISINFORMATION

32 Safety Tech Innovation Network, 'Trust, Safety and the Digital Economy, 2022. https://www.safetytechnetwork.org.uk/trust-safety-and-the-digital-economy/; London School of Economics, 'The Cost of Lies. Assessing the human and financial impact of COVID-19 related online misinformation on the UK,' 2020, https://londoneconomics.co.uk/wp-content/uploads/2021/01/The-Cost-of-Lies_clean_2.2.21.pdf

2 Harms to human rights and democracy

Numerous studies have established that:

- False and misleading content on the platforms is widely perceived as a major threat to democracy. Among various studies,[33] one survey of 154195 regular internet users in 142 countries showed that almost 60% worry about misinformation, and young and low-income groups are most likely to be concerned.[34] According to Datareportal, 43% of young people agree that algorithms that determine what they see in their online content feeds have a negative impact on their media diet.[35]
- A meta-analysis of almost 500 relevant studies about digital media, trust and polarisation, noted some ambiguity but also pinpointed a correlation: "Results also suggest that digital media use is associated with increases in hate, populism, and polarisation".[36] Echo-chambers and in-group reinforcement are regularly recorded.[37] Political radicalisation, including via YouTube and through Whatsapp groups, is also a phenomenon identified by a number of researchers.[38] A long list of other harms to various human rights have come to light, including many findings based on company documents leaked by whistle blowers.[39]
- In the words of Nobel Prize-winning journalist Maria Rezza, corroborated by much research: "Lies laced with anger and hate spread faster and further than facts".[40] Such falsehoods penetrate widely on both niche and broader platforms, and are linked to growing public distrust in science and elections.[41] Researchers have also found a correlation between social media use and decreased resilience to misinformation.[42]

---

33 Chris Tenove, 'Protéger la démocratie de la désinformation: Menaces normatives et réponses politiques', 23 July 2020, https://crtc.gc.ca/fra/acrtc/prx/2020tenove.htm; Linda Sanchez, 'RENFORCER LA RÉSILIENCE DÉMOCRATIQUE DE L'ALLIANCE FACE À LA DÉSINFORMATION ET LA PROPAGANDE' 2 (10 October 2021).

34 Aleksi Knuutila, Lisa-Maria Neudert, and Philip N. Howard, 'Who Is Afraid of Fake News? Modeling Risk Perceptions of Misinformation in 142 Countries', *Harvard Kennedy School Misinformation Review*, 12 April 2022, https://doi.org/10.37016/mr-2020-97; Martin N. Ndlela and Winston Mano, eds., *Social Media and Elections in Africa, Volume 1: Theoretical Perspectives and Election Campaigns* (Cham: Springer International Publishing, 2020), https://doi.org/10.1007/978-3-030-30553-6.

35 Kemp, 'The Global State of Digital in October 2022 — DataReportal – Global Digital Insights'.

36 Philipp Lorenz-Spreen et al., 'A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy', preprint (SocArXiv, 22 November 2021), https://doi.org/10.31235/osf.io/p3z9v.preprint (SocArXiv, 22 November 2021

37 Magdalena Wojcieszak et al., 'Most Users Do Not Follow Political Elites on Twitter; Those Who Do Show Overwhelming Preferences for Ideological Congruity', *Science Advances* 8, no. 39 (30 September 2022): eabn9418, https://doi.org/10.1126/sciadv.abn9418; Jakob Ohme, 'Algorithmic Social Media Use and Its Relationship to Attitude Reinforcement and Issue-Specific Political Participation – The Case of the 2015 European Immigration Movements', *Journal of Information Technology & Politics* 18 (2 September 2020): 1–18, https://doi.org/10.1080/19331681.2020.1805085; Tech Transparency Project, 'Facebook Profits from White Supremacist Groups'; Megan A. Brown et al., 'Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users', SSRN Scholarly Paper (Rochester, NY, 11 May 2022), https://doi.org/10.2139/ssrn.4114905; Matteo Cinelli et al., 'Dynamics of Online Hate and Misinformation', *Scientific Reports* 11, no. 1 (11 November 2021): 22083, https://doi.org/10.1038/s41598-021-01487-w.\\uc0\\u8216{} Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users\\uc0\\u8217{}, SSRN Scholarly Paper (Rochester, NY, 11 May 2022

38 Rafael Evangelista and Fernanda Bruno, 'WhatsApp and Political Instability in Brazil: Targeted Messages and Political Radicalisation', *Internet Policy Review* 8, no. 4 (31 December 2019), https://doi.org/10.14763/2019.4.1434; John D Gallacher and Jonathan Bright, 'Hate Contagion: Measuring the Spread and Trajectory of Hate on Social Media', preprint (PsyArXiv, 16 February 2021), https://doi.org/10.31234/osf.io/b9qhd; Justin Hendrix, 'Can Big Tech Platforms Operate Responsibly on a Global Scale?', Tech Policy Press, 18 September 2022, https://techpolicy.press/can-big-tech-platforms-operate-responsibly-on-a-global-scale/.

39 Accountable Tech, 'Ban Surveillance Advertising', https://accountabletech.org/campaign/ban-surveillance-advertising/.

40 William T Adler and Dhanaraj Thakur, 'A Lie Can Travel: Election Disinformation in the United States, Brazil, and France', *Center for Democracy and Technology* (blog), 2021, https://cdt.org/insights/cdt-and-kas-report-a-lie-can-travel-election-disinformation-in-the-united-states-brazil-and-france/; Soroush Vosoughi, Deb Roy, and Sinan Aral, 'The Spread of True and False News Online', *Science* 359, no. 6380 (9 March 2018): 1146–51, https://doi.org/10.1126/science.aap9559; Matthew Shaer, 'What Emotion Goes Viral the Fastest?', Smithsonian Magazine, 2014, https://www.smithsonianmag.com/science-nature/what-emotion-goes-viral-fastest-180950182/; Lennart Maschmeyer, 'Désinformation en ligne: le cas de l'Ukraine', application/pdf, February 2021, 4 p., https://doi.org/10.3929/ETHZ-B-000465902.

41 Justin Hendrix, 'YouTube and the "Big Lie": Research Shows Cause for Concern', Tech Policy Press, 2 September 2022, https://techpolicy.press/youtube-and-the-big-lie-research-shows-cause-for-concern/.

42 Shelley Boulianne, Chris Tenove, and Jordan Buffie, 'Complicating the Resilience Model: A Four-Country Study About Misinformation', *Media and Communication* 10, no. 3 (31 August 2022): 169–82, https://doi.org/10.17645/mac.v10i3.5346; Brian Owens, 'Social-Media Platforms Failing to Tackle Abuse of Scientists', *Nature* 602, no. 7896 (10 February 2022): 197–197, https://doi.org/10.1038/d41586-022-00207-2.

- Notwithstanding denials by some platform representatives,[43] algorithmic rankings and recommendations are broadly recognised as being factors which intensify content that is characterised by incivility and polarisation.[44] It would be unscientific to attribute any societal impact exclusively to social media platforms, and many studies have recognised other drivers of political polarisation.[45] However, even Meta boss Mark Zuckerberg has stated: "One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. … At scale it can undermine the quality of public discourse and lead to polarisation…."[46] In this context, researcher Jonathan Stray argues that "Polarisation is implicated in the erosion of democracy and the progression to violence, which makes the polarisation properties of large algorithmic content selection systems (recommender systems) a matter of concern for peace and security".[47]
- Content on the platforms is widely seen as not just fuelling divisions but helping to fuel gross human rights violations at scale, such as through incitement to violence, and the organisation thereof.[48]

Assessing the factors that account for online content that potentially or actually harms human rights is essential for exploring what governance changes could work best to address the problems without compounding the situation.

"One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. … At scale it can undermine the quality of public discourse and lead to polarisation…."

43 Nick Clegg, 'You and the Algorithm: It Takes Two to Tango', *Medium* (blog), 31 March 2021, https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2.

44 Yingying Chen and Luping Wang, 'Misleading Political Advertising Fuels Incivility Online: A Social Network Analysis of 2020 U.S. Presidential Election Campaign Video Comments on YouTube', *Computers in Human Behavior* 131 (June 2022): 107202, https://doi.org/10.1016/j.chb.2022.107202; Michael Henry Yusingco, 'Social Media, Disinformation, and the 2022 BARMM Parliamentary Elections', *SSRN Electronic Journal*, 2021, https://doi.org/10.2139/ssrn.3763264; Petter Törnberg, 'How Digital Media Drive Affective Polarization through Partisan Sorting', *Proceedings of the National Academy of Sciences* 119, no. 42 (18 October 2022): e2207159119, https://doi.org/10.1073/pnas.2207159119; Adi Cohen, 'How Social Media Fanned the Flames of Israeli-Palestinian Violence', Tech Policy Press, 3 June 2021, https://techpolicy.press/how-social-media-fanned-the-flames-of-israeli-palestinian-violence/.

45 Paul M. Barrett, 'Spreading the Big Lie: How Social Media Sites Have Amplified False Claims of U.S. Election Fraud', 2022. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/6321e2ca392ecd06e5d60c4c/1663165130817/NYU+Stern+Center+report-+Spreading+the+Big+Lie_FINAL.pdf

46 Mark Zuckerberg, 'Blueprint for Content Governance and Enforcement', 2021, https://www.facebook.com/notes/751449002072082/

47 Jonathan Stray, 'Designing Recommender Systems to Depolarize' (arXiv, 10 July 2021), https://doi.org/10.48550/arXiv.2107.04953. which makes the polarization properties of large algorithmic content selection systems (recommender systems

48 Sibiri Yeo, 'TIC ET VIOLENCES ÉLECTORALES EN CÔTE D'IVOIRE', 2021; Giovanni De Gregorio and Nicole Stremlau, 'Internet Shutdowns in Africa | Internet Shutdowns and the Limits of Law', *International Journal of Communication* 14, no. 0 (13 August 2020): 20; Maria Pawelec, 'Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions', *Digital Society* 1, no. 2 (September 2022): 19, https://doi.org/10.1007/s44206-022-00010-6; Amnesty International, 'Myanmar: The Social Atrocity: Meta and the Right to Remedy for the Rohingya', Amnesty International, 29 September 2022, https://www.amnesty.org/en/documents/asa16/5933/2022/en/; 'Facebook Accused by Survivors of Letting Activists Incite Ethnic Massacres with Hate and Misinformation in Ethiopia', The Bureau of Investigative Journalism (en-GB), accessed 19 December 2022, https://www.thebureauinvestigates.com/stories/2022-02-20/facebook-accused-of-letting-activists-incite-ethnic-massacres-with-hate-and-misinformation-by-survivors-in-ethiopia; 'Hate Speech in Myanmar Continues to Thrive on Facebook', AP NEWS, 18 November 2021, https://apnews.com/article/technology-business-middle-east-religion-europe-a38da3ccd40ffae7e4caa450c374f796; Jasper Jackson, Mark Townsend, and Lucy Kassa, 'Facebook "Lets Vigilantes in Ethiopia Incite Ethnic Killing"', *The Observer*, 20 February 2022, sec. Technology, https://www.theguardian.com/technology/2022/feb/20/facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing; Karen Hao, 'How Facebook and Google Fund Global Misinformation', 2021, https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/; European Institute of Peace, 'Fake News Misinformation and Hate Speech in Ethiopia: A Vulnerability Assessment | EIP', 2021, https://www.eip.org/publication/fake-news-misinformation-and-hate-speech-in-ethiopia-a-vulnerability-assessment/.{\\i{}Digital Society} 1, no. 2 (September 2022

3 Unpacking the causes

The underlying roots of many ills on the platforms can be grouped into five fundamental categories - 1. Attention-economics and micro-targeted advertising, 2. Automated advertising, 3. External manipulation, 4. Company spending priorities, and 5. Stakeholder knowledge deficits. The additional contribution of the platforms' own policies and related implementation is dissected in Part 2 in this series.

'ATTENTION ECONOMICS' AND MICRO-TARGETED ADVERTISING

A major driver of problematic content online is the targeted advertising logic adopted by many of the commercial platforms whether these entities are broadly targeted or aimed at niche ideological or other markets. For social media and video-sharing services, this logic results in algorithms being customised to cultivate attention and engagement. The more attention, the more the opportunity for platforms to reap audience data and to show customised advertising. This logic then functions to foster the production and spread of emotive discourse[49] that is also often infused with falsehoods[50], conspiracy theories and hate speech.[51] While the term "users" implies that people who avail themselves of the platforms' services are in the driving seat, the actual business model at work means these individuals themselves are being used. The system is designed to convert them into distinctive data "commodities" sold to paying customers as opportunities to do precision targeting with advertisements.[52] The logic of cultivating attention influences what is prioritised in people's content and advertising feeds, as well as what they are recommended.

An illustration of the significance of attention economics is the case of YouTube where automated recommendations systems aimed at keeping viewers watching have worked to encourage users down a rabbit-hole of ever more extreme content. The journalist Kevin Roose has uncovered how these processes can result in radicalisation towards extremist violence.[53] Algorithms assessed what users were watching and directed them to further and more intense offerings on the same theme.[54] Similarly demonstrating the equation whereby "enragement=engagement", Instagram's recommendation algorithm has led users interacting with anti-vaccination misinformation onwards towards posts containing anti-semitism and election conspiracy theories.[55] Roose has further revealed that Facebook's top ten posts for engagement daily tend to be dominated by angry voices, often from the extreme right wing[56] – a phenomenon that becomes algorithmically self-perpetuating as new people decide to follow "trending topics". As concerns Twitter, this platform has been described as a "digital outrage engine" in response to how its CEO has bluntly explained how its content feed works.[57]

---

49 Carlos Carrasco-Farré, 'The Fingerprints of Misinformation: How Deceptive Content Differs from Reliable Sources in Terms of Cognitive Effort and Appeal to Emotions', *Humanities and Social Sciences Communications* 9, no. 1 (9 May 2022): 1–18, https://doi.org/10.1057/s41599-022-01174-9.fake news, junk science, or rumors among others. However, most of the existing research does not account for these differences. This paper explores the characteristics of misinformation content compared to factual news—the "fingerprints of misinformation"—using 92,112 news articles classified into several categories: clickbait, conspiracy theories, fake news, hate speech, junk science, and rumors. These misinformation categories are compared with factual news measuring the cognitive effort needed to process the content (grammar and lexical complexity

50 Ullrich K. H. Ecker et al, 'The psychological drivers of misinformation belief and its resistance to correction', 2022, https://www.nature.com/articles/s44159-021-00006-y.pdf

51 Lan Ha, Timothy Graham, and Joanne Gray, 'Where Conspiracy Theories Flourish: A Study of YouTube Comments and Bill Gates Conspiracy Theories', *Harvard Kennedy School Misinformation Review*, 5 October 2022, https://doi.org/10.37016/mr-2020-107.

52 Shoshana Zuboff, '*The age of surveillance capitalism. The fight for a human future at the new frontier of power*', Public Affairs.

53 Kevin Roose, 'The making of a YouTube radical', 2019, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

54 Kevin Roose, 'The Making of a YouTube Radical', *The New York Times*, 8 June 2019, sec. Technology, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html; Bharat Sharma, 'YouTube Says It's Unlike Other Platforms, Recommends Less Than 1% "Borderline Content"', IndiaTimes, 20 October 2021, https://www.indiatimes.com/technology/news/youtube-borderline-content-recommendation-552155.html.

55 CCDH, 'Failure to Protect. Social Media Platforms Are Failing to Act on Anti-Muslim Hate', 2022.

56 Kevin Roose, 'Facebook's Top 10 (@FacebooksTop10) / Twitter', Twitter, 2022, https://twitter.com/FacebooksTop10.

57 Carl Bergstrom. 'Elon Musk and Twitter as a digital outrage engine', 2023, https://post.news/article/2KTDq7DXiMkp5J31wEMwMXB33tX

Research into this company has also found that in 6 of 7 countries studied, the algorithm has disproportionately amplified the mainstream political right, as well as right-leaning news sources.[58]

It has become a debate as to whether specifically algorithmic recommendation of content constitutes an editorial intervention by platforms, rather than simply being a neutral mediation of third-party content. This issue has underpinned a case against Google to the effect that the curation of recommendations voids the company's legal immunity for liability as regards third party content.[59] It may be that a search query which yields list of links including illegal content, is less vulnerable in legal terms than would be a case where automated recommendations of the same content get proactively proposed to users who have not solicited them but have been auto-targeted on a certain logic. Apart from liability issues, there is some momentum for the rationales behind recommendations of potentially harmful content to be considered in various regulatory arrangements (including better solo-regulation measures), and that recommendations that are particularly driven by attention economics merit particular attention.

In another illustration of attention economics driving content, an algorithmic change by Facebook aimed at promoting content from family and friends (and demoting public content like posts from news publishers[60]), had the outcome of foregrounding the angriest expression coming from this constituency.[61] "Facebook works on feelings, not facts", is the insight of researcher Siva Vaidhyanathan.[62] Conversely, algorithms can also be altered to work in the other direction. Faced with negative publicity, and to limit the reach of misinformation the company deemed to be harmful, YouTube tweaked its recommendation algorithms. The company said the result led to users spending 70 percent less time watching what it called "borderline" videos.[63] The case shows that attention economics can be reined in, although at present this option is left to companies' own choices.

Beyond promoting anger and division, platform recommendation engines stand accused of escalating risks to children's rights. One study of teenager sign-ups to TikTok showed that suicide content was recommended within 2.6 minutes, and eating disorder content within 8 minutes, with content about the latter topic reaching over 13.2 billion views.[64]

FUEL FOR RACISM: Research giving insight into the workings of attention economics uncovered 80 white supremacist groups on Facebook, associated with at least 119 Facebook Pages and 20 Facebook Groups. After "liking" the 119 pages, the researchers found that in 58% of the cases Facebook pointed them onward towards other extremist or hateful content. Searches for the names of white supremacist groups on Facebook often produced search results with adverts, meaning that Facebook made money accordingly. Redirecting users who search for terms associated with hate groups to organisations that promote tolerance only worked in 14% of 226 searches for white supremacist organisations.[65]

58 Ferenc Huszár. Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer and Moritz Hardt, 'Algorithmic Amplification of Politics on Twitter', 2021, https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf

59 'Reynaldo Gonzalez, et al., Petitioners v.Google LLC', 2022, https://www.supremecourt.gov/docket/docketfiles/html/public/21-1333.html.

60 Adam Mosseri, 'Bringing People Closer Together | Meta', 2018, https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/.

61 Washington Post, 'Why Facebook Won't Let You Control Your Own News Feed', News, Washington Post, 13 November 2021, https://www.washingtonpost.com/technology/2021/11/13/facebook-news-feed-algorithm-how-to-turn-it-off/; The Journal, 'The Facebook Files, Part 4: The Outrage Algorithm - The Journal. - WSJ Podcasts', WSJ, 2021, https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f.

62 Siva Vaidhyanathan, 'Making Sense of the Facebook Menace', *The New Republic*, 5 January 2021, https://newrepublic.com/article/160661/facebook-menace-making-platform-safe-democracy.

63 Greg Bensinger, 'YouTube Says Viewers Are Spending Less Time Watching Conspiracy Theory Videos. But Many Still Do.', *Washington Post*, 29 January 2020, https://www.washingtonpost.com/technology/2019/12/03/youtube-says-viewers-are-spending-less-time-watching-conspiracy-videos-many-still-do/; The YouTube Team, 'Managing Harmful Conspiracy Theories on YouTube', blog.youtube, 2020, https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/.

64 Center for Countering Digital Hate, 'Deadly by design', 2023, https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf

65 Tech Transparency Project, 'Facebook Profits from White Supremacist Groups', https://www.techtransparencyproject.org/articles/facebook-profits-white-supremacist-groups.

In the face of public criticism of attention-fostering algorithms, some companies have begun to offer the option for people to receive a purely chronological feed of content. However, as several giant platforms move towards a default feed based on a "TikTok-ified" algorithm of endless short videos, it is evident that this trend could further intensify more emotive and frivolous offerings at the expense of informative and educative content, and increase video as a proportion of content that is more difficult to moderate than text.[66] Relevant to this scenario is a study of TikTok in 2021 which found that white supremacists and terrorist groups used music and video effects to promote their cause.[67]

Despite a suite of interventions by the platforms[68], and notwithstanding many appeals to them to review their business modus operandi[69], the fundamental model of attention economics persists.[70] Although platform companies also acquire people's personal data from external brokers (who aggregate, exploit and sell data from other online sources), their ability to directly collect and mine first-party data is a major component of the attention economics model. As may be expected, the companies are reluctant to voluntarily limit their ability to retain attention as well as limit the opportunities to sell advertising slots that depend on the screen time which people need to be nudged to bestow on the platform concerned.

What all this draws attention to is the distinction between prevalent logics in curation and moderation.[71] The architecture of many platforms is fundamentally designed to curate content and communications (including ranking and recommendations) to achieve optimum attention as a means to increasing advertising revenues. It follows that many active moderation efforts (whether automated or not) are generally layered on top of this, and that they therefore represent attempts to do ex post facto damage-control of negative effects that are the consequences of the particular configurations of curation.

The influence of curation driven by attention economics is evident in several reports. For example, Facebook uses as many as 10 000 data points (from various sources including through their own tracking user activity on the app as well as scraped from elsewhere on the web[72]). Platforms accumulate potentially sensitive user data such as location, device used, IP address, search history, contacts, message content and what content is being consumed and shared, for how long, and at what time of day, as well as biometric data. These data are dynamically assessed to shape the news feed selection and prioritisation of what individual users will receive.[73] The same system, which has been described as a "Digital Influence Machine"[74] is what powers the targeting of advertising content to these users.[75] Correspondingly, companies have been found to be gravely lacking in their disclosure about how users' online content (including advertising) is curated.[76]

As may be expected, the companies are reluctant to voluntarily limit their ability to retain attention as well as limit the opportunities to sell advertising slots that depend on the screen time which people need to be nudged to bestow on the platform concerned.

66 Abby Ohlheiser, 'The Most Popular Content on Facebook Belongs in the Garbage | MIT Technology Review', 25 August 2022, https://www.technologyreview.com/2022/08/25/1058662/facebook-widely-viewed-content-spam-memes/.

67 Ciaran O'Connor, 'The Extreme Right on Twitch', 2021; Álvarez Ugarte, 'Fake News on the Internet: Actions and Reactions of Three Platforms. Presentation of CELE before the UN Special Rapporteur for Freedom of Opinion and Expression. - CELE Fake News on the Internet: Actions and Reactions from Three Platforms', *CELE* (blog), 18 February 2021, https://observatoriolegislativocele.com/en/fake-news-on-the-internet-actions-and-reactions-of-three-platforms-presentation-of-the-cele-before-the-special-rapporteur-for-freedom-of-opinion-and-expression-of-the-un/.

68 Ramiro Álvarez Ugarte and and Agustina Del Campo, 'Fake News on the Internet: actions and reactions of three platforms', February 2021, https://www.palermo.edu/Archivos_content/2021/cele/papers/Fake-news-on-the-Internet-2021.pdf

69 Irene Khan, 'OHCHR | Special Rapporteur on Freedom of Opinion and Expression', OHCHR, 2021, https://www.ohchr.org/en/special-procedures/sr-freedom-of-opinion-and-expression.

70 Cai Hui Lien, James Lee, and Edson C. Tandoc, 'Facing Fakes: Understanding Tech Platforms' Responses to Online Falsehoods', *Digital Journalism* 10, no. 5 (28 May 2022): 761–80, https://doi.org/10.1080/21670811.2021.1982398.

71 Eliška Pírková and Javier Pallero, '26 recommendations on content governance: a guide for lawmakers, regulators, and company policy makers', AccessNow, 2020, https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf

72 Marissa Newman, 'Meta Was Scraping Sites for Years While Fighting the Practice', 2023 https://www.bloomberg.com/news/articles/2023-02-02/meta-was-scraping-sites-for-years-while-fighting-the-practice?cmpid=socialflow-twitter-business&leadSource=uverify+wall

73 Washington Post, 'Why Facebook Won't Let You Control Your Own News Feed', https://www.washingtonpost.com/technology/2021/11/13/facebook-news-feed-algorithm-how-to-turn-it-off/.

74 Anthony Nadler, Matthew Crain, and Joan Donovan, 'Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech' (Data & Society, 2018), https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf.

75 Accountable Tech, 'Ban Surveillance Advertising'.

76 Ranking Digital Rights, 'The 2022 Ranking Digital Rights Big Tech Scorecard', https://rankingdigitalrights.org/rankings-report-cards/.

AUTOMATED ADVERTISING

A second factor that explains the phenomenon of harmful content online, is the reliance by many platforms on automated (and barely regulated) advertising. These systems rely in turn on assembly and ownership of massive data holdings that can be subjected to real-time analysis.[77] This mining of data for advertising applies even to services that do not depend on "attention economics", such as search engines and search functionalities on other platforms. These 'search' their 'users' by means of monitoring what is being searched for, and they use this data to add to the advertising targeting. The common aim is to discern data patterns in order to sell advertising opportunities with pinpoint customisation aimed at nudging respondents on the basis of their profiles, interests and behaviours. According to Shoshana Zuboff, the trade amounts to being one of "behavioural futures". [78]

Although companies usually have specific content rules for advertising, the efficacy of these is weakened by "programmatic" (meaning "computational") mechanisms for buying, selling and placing this kind of content. The ad transactions occur in real-time automated auctions within opaque value chains where neither advertisers nor publishers are aware of all the processes and fees deducted along the way.[79] Even Whatsapp, which generally does not sell advertising, produces data such as phone numbers and device information which Meta then uses to computationally match accounts with its Facebook and Instagram.[80]

Running on an automated basis that privileges transactions particularly enables micro-targeting with potentially harmful paid-for content. The system also facilitates advertising revenues to flow not only to the platform concerned but also sometimes to the producers of such problematic content, including via link-based traffic sent to their websites.[81] Illustrations all this are:

- Ad-tech companies (in which Google and Facebook provide leading services) have placed advertisements on 20 000 sites flagged for disinformation.[82] Google says it has made a deliberate change to its search algorithms with significantly reduced the discoverability of "junk news",[83] but according to the company NewsGuard (an entity which itself is accused of legitimating "junk news" sites for advertisers[84]) the current configuration of ad-tech meant that over 1,000 brands ended up with approximately 8,776 advertisements containing US electoral misinformation being placed on 160 sites during several months prior to mid-January 2021.[85]
- Spreaders of climate disinformation, dubbed the "Toxic Ten" by researchers studying this phenomenon, notched up 186 million followers on social media and were able to generate up to $5.3 million in Google Ads revenue during six months.[86]
- Another study found that 48% of all advertising traffic on a number of "fake" news sites is served by Google's own automated buying and placement systems, while 32% of adverts served on other "low quality" sites also comes via the same company.[87]

77 Tom Dobber, Ronan Ó Fathaigh, and Frederik J. Zuiderveen Borgesius, 'The Regulation of Online Political Micro-Targeting in Europe', *Internet Policy Review* 8, no. 4 (31 December 2019), https://doi.org/10.14763/2019.4.1440; Somya Mehta and Kristofer Erickson, 'Can Online Political Targeting Be Rendered Transparent? Prospects for Campaign Oversight Using the Facebook Ad Library', *Internet Policy Review* 11, no. 1 (31 March 2022), https://policyreview.info/articles/analysis/can-online-political-targeting-be-rendered-transparent-prospects-campaign.

78 Shoshana Zuboff, 'The age of surveillance capitalism'

79 See: Rebecca Frank, 'Ad Tech: It's Worse Than We Thought', 2022, https://www.newsmediaalliance.org/ad-tech-its-worse-than-we-thought/; European Publishers Council, 'Our competition complaint', 2022, https://www.epceurope.eu/complaint; US Department of Justice, 'Justice Department Sues Google for Monopolizing Digital Advertising Technologies', 2023, https://www.justice.gov/opa/video/justice-department-sues-google-monopolizing-digital-advertising-technologies

80 Meta Platforms, Inc., 'ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934For the Fiscal Year Ended December 31, 2021'.

81 Center for Countering Digital Hate, 'The Disinformation Dozen', Center for Countering Digital Hate | CCDH, accessed 20 December 2022, https://counterhate.com/research/the-disinformation-dozen/.

82 GDI, 'The Quarter Billion Dollar Question: How Is Disinformation Gaming Ad Tech?'

83 Samantha Bradshaw, 'Disinformation Optimised: Gaming Search Engine Algorithms to Amplify Junk News', *Internet Policy Review* 8, no. 4 (31 December 2019), https://doi.org/10.14763/2019.4.1442.

84 Check My Ads Institute, 'How NewsGuard Boosts the American Disinformation Economy', Check My Ads Institute, 11 August 2022, https://checkmyads.org/branded/how-newsguard-boosts-the-american-disinformation-economy/.

85 Aspen Institute, 'Commission on Information Disorders Final Report', 2021.

86 Center for Countering Digital Hate, 'Malgorithm. How Instagram's Algorithm Publishes Misinformation and Hate During the Pandemic', 2021.

87 Lia Bozarth and Ceren Budak, 'Market Forces: Quantifying the Role of Top Credible Ad Servers in the Fake News Ecosystem', *Proceedings of the International AAAI Conference on Web and Social Media* 15 (22 May 2021): 83–94, https://doi.org/10.1609/icwsm.v15i1.18043.

- The NGO "CheckMyAds" has found that 9 out of 10 sellers on Google Ads are untraceable, allowing actors to monetise their websites without even transparency to the advertisers.[88]
- Propublica scanned more than seven million website domains for Google advertising activity and found that the vast majority of advertising selling partners were hidden, which allowed for the promotion of content piracy, pornography, fraud and disinformation through ads automatically placed on unwitting publishers' sites.[89]

Among the problems of automated advertising, it has also been reported that the "gaming" of ad-tech bargaining systems with emotive content reduced costs and expanded reach.[90] Test adverts promoting drug abuse, anorexia and gambling have been approved for show in less than an hour.[91] Stealth political adverts are able to bypass company systems.[92] Numerous instances are recorded of advertisements that infringe platform policy.[93] According to monitoring group Ranking Digital Rights, only 5 of 14 major platforms publish any data about actions taken to restrict advertising that violates their policies.[94] Further, a study of 11 countries and the EU found that regulatory efforts did not adequately address the financial incentives or amplification of disinformation.[95]

For its part, Google (speaking mainly for its video sharing and search platforms) says "our advertising products are among the most [policy] restrictive, as we do not want to profit from those who create hate or harmful experiences."[96] The company adds that in 2019 it removed more than 2.7 billion "bad ads"; took action against almost one million advertiser accounts; terminated over 1.2 million publisher accounts; and removed ads from over 21 million web pages. These figures prompt the question as to whether this equates to a Sisyphus-like quest to counter problems that are inherent in the ad-tech business as currently configured.

In the case of TikTok, attention economics combined with the elimination of conscious decision-making from users[97], along with micro-targeting capacity around both organic content and paid advertising, entails seamless continuity and reduced visual distinction between the two formats.[98] 'Native advertising' formats in particular are intended to be indistinguishable from other content on the feed.[99]

88 Check My Ads Institute, 'Google Ads Has Become a Massive Dark Money Operation', Check My Ads Institute, 18 October 2022, https://checkmyads.org/branded/google-ads-has-become-a-massive-dark-money-operation/.

89 Craig Silverman and Ruth Talbot, 'Porn, Piracy, Fraud: What Lurks Inside Google's Black Box Ad Empire', Propublica, 21 December, 2022, https://www.propublica.org/article/google-display-ads-piracy-porn-fraud

90 Kumar Sambhav and Nayantara Ranganathan, 'Facebook Charged BJP Less for India Election Ads than Others', 2022, https://www.aljazeera.com/economy/2022/3/16/facebook-charged-bjp-lower-rates-for-india-polls-ads-than-others; Nayantara Ranganathan and Kumar Sambhav, 'What Helps India's BJP Get Lower Facebook Rates? Divisive Content', 2022, https://www.aljazeera.com/economy/2022/3/17/facebook-algorithm-favours-polarising-politics-helps-bjp.

91 Tech Transparency Project, 'Facebook Profits from White Supremacist Groups'.

92 Laura Edelson, Tobias Lauinger, and Damon McCoy, 'A Security Analysis of the Facebook Ad Library', in *2020 IEEE Symposium on Security and Privacy (SP)* (2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA: IEEE, 2020), 661–78, https://doi.org/10.1109/SP40000.2020.00084.

93 GDI, 'Ad Tech Policy and Enforcement Gaps: Challenges and Solutions', 2022.

94 Ranking Digital Rights, 'The 2022 Ranking Digital Rights Big Tech Scorecard', Ranking Digital Rights, accessed 18 December 2022, https://rankingdigitalrights.org/bts22/indicators/F4b.

95 GDI, 'Disrupting Disinformation: A Global Snapshot of Government Initiatives', 2021.

96 Google Youtube, 'Information Quality & Content Moderation'.

97 Arvind Narayanan, 'TikTok's Secret Sauce', 2022, https://knightcolumbia.org/blog/tiktoks-secret-sauce

98 See Dare Obasanjo, 2022. https://mas.to/@carnage4life/109688112381311 94; Tor Marom, 'TikTok Ads Regularly Bring In 2.5X ROAS — But That's Not the Only Reason to Try Them', 2022, https://marketerhire.com/blog/tiktok-ads3;

99 John Williams, 'Paid and Organic are Indistinguishable', 2018, https://www.linkedin.com/pulse/paid-organic-indistinguishable-john-williams-

According to researchers Nicholas Carah and Sven Brodmerkel, for whom the focus should be deeper than the particular content carried in adverts, "the value proposition of a digital platform isn't its ability to 'place' the ad in front of the right person at the right time, in so much as it is to gradually tune and tweak its capacity to nudge, move, engage a consumer".[100]

Overall, and relevant to both advertisements and non-paid-for content, the significance is framed by researcher Kate Jones as follows: "The rights to freedom of thought and opinion are critical to delimiting the appropriate boundary between legitimate influence and illegitimate manipulation." She argues that when digital platforms exploit decision-making biases in prioritising bad news and divisive, emotion-arousing content, they may be intruding on these aspects of personal autonomy as enshrined in the right to freedom of opinion. For her, "States and digital platforms should consider structural changes to digital platforms to ensure that methods of online political discourse respect personal agency and prevent the use of sophisticated manipulative techniques."[101] Complementing this view, the World Economic Forum has proposed standards for platforms around recommendations which could help distinguish between helpful personalised suggestions and undue influence.[102]

"The rights to freedom of thought and opinion are critical to delimiting the appropriate boundary between legitimate influence and illegitimate manipulation."

EXTERNAL MANIPULATION

Another cause of online content that is potentially harmful to human rights is the exploitation of the platforms by external actors seeking manipulative political and/or financial gain through the use of communications, content activities and advertising.[103] This raises questions of how these actors are able to exploit the platforms to do so, and how the platforms themselves deal with the problem.

That there is indeed a major problem with manipulation is increasingly evident. Numerous cases exist of outsourced trolling and exploitation of gender-based hate to boost expression that may harm human rights.[104] The scale, as companies like to point out, is almost inconceivable - especially if account is taken

100 Nicholas Carah and Sven Brodmerkel, 'Regulating Platforms' Algorithmic Brand Culture: The Instructive Case of Alcohol Marketers on Social Media', in *Digital Platform Regulation: Global Perspectives on Internet Governance*, ed. Terry Flew and Fiona R. Martin, Palgrave Global Media Policy and Business (Cham: Springer International Publishing, 2022), 111–30, https://doi.org/10.1007/978-3-030-95220-4_6.and argues that understanding the engineering, operation and consequences of platforms' data-driven, participatory and opaque advertising model is fundamental to addressing larger questions of platform regulation in the public interest. It suggests that through the case of alcohol marketing we can understand and assess many of the novel regulatory challenges posed by the advertising model of digital platform companies. Thus, in this chapter we appraise some of the existing alcohol industry and platform approaches to self-regulation and suggest some principles for regulating marketing that is data-driven, participatory and opaque, and connect these to larger debates about the future regulation of platforms. The critical assessment of the novel ways in which platform marketing integrates participatory forms of audience engagement with the prospecting, segmentation and targeting of consumers is crucial for developing an accountable regulatory regime that allows for effective governance of the commercial activities of marketers and brands on platforms.","collection-title":"Palgrave Global Media Policy and Business","container-title":"Digital Platform Regulation: Global Perspectives on Internet Governance","event-place":"Cham","ISBN":"978-3-030-95220-4","language":"en","note":"DOI: 10.1007/978-3-030-95220-4_6","page":"111-130","publisher":"Springer International Publishing","publisher-place":"Cham","source":"Springer Link","title":"Regulating Platforms' Algorithmic Brand Culture: The Instructive Case of Alcohol Marketers on Social Media","title-short":"Regulating Platforms' Algorithmic Brand Culture","URL":"https://doi.org/10.1007/978-3-030-95220-4_6","author":[{"family":"Carah","given":"Nicholas"},{"family":"Brodmerkel","given":"Sven"}],"editor":[{"family":"Flew","given":"Terry"},{"family":"Martin","given":"Fiona R."}],"accessed":{"date-parts":[["2022",12,18]]},"issued":{"date-parts":[["2022"]]}}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"}

101 See also: Eliška Pírková and Javier Pallero, '26 recommendations on content governance: a guide for lawmakers, regulators, and company policy makers'

102 World Economic Forum. 'White paper: Advancing Digital Safety: A Framework to Align Global Action', 2021, https://www3.weforum.org/docs/WEF_Advancing_Digital_Safety_A_Framework_to_Align_Global_Action_2021.pdf

103 Rowan Philp, 'Comment Enquêter Sur La Désinformation En Période d'élections', 2022, https://gijn.org/2022/05/23/francais-election-desinformation/; Benjamin Strick, 'Investigating Information Operations in West Papua: A Digital Forensic Case Study of Cross-Platform Network Analysis', bellingcat, 11 October 2019, https://www.bellingcat.com/news/rest-of-world/2019/10/11/investigating-information-operations-in-west-papua-a-digital-forensic-case-study-of-cross-platform-network-analysis/; Ben Nimmo, 'Meta's Adversarial Threat Report, Second Quarter 2022 | Meta', 2022, https://about.fb.com/news/2022/08/metas-adversarial-threat-report-q2-2022/; Barrett, 'Spreading The Big Lie: How Social Media Sites Have Amplified False Claims of U.S. Election Fraud'; Hendrix, 'Can Big Tech Platforms Operate Responsibly on a Global Scale?'; Jack Stubbs and Shawn Eib, 'Coordinated Inauthentic Bee-havior', Graphika, accessed 19 December 2022, https://graphika.com/reports/coordinated-inauthentic-bee-havior; Renaud De la Brosse, Jean-François Furnemont, and Abdourahamane Ousmane, 'LA LUTTE CONTRE LA DÉSINFORMATION DANS LES POLITIQUES PUBLIQUES FRANCOPHONES', n.d.

104 Shelby Grossman, Sean Gallagher, and Ada Johnson-Kanu, '#ZakzakyLifeMatters: An Investigation into a Facebook Operation Linked to the Islamic Movement in Nigeria', n.d.; Shelby Grossman et al., 'Reply-Guys Go Hunting: An Investigation into a U.S. Astroturfing Operation on Facebook, Twitter, and Instagram', 2020; Renee DiResta et al., 'Mind Farce: An Investigation into an Inauthentic Facebook and Instagram Network Linked to an Israeli Public Relations Firm (TAKEDOWN)', 4 August 2022, https://fsi.stanford.edu/publication/mind-farce-investigation-inauthentic-facebook-and-instagram-network-linked-israeli; Samantha Bradshaw, 'The Gender Dimensions of Foreign Influence Operations', 2021; Shelby Grossman et al., 'My Heart Loves the Army: An Investigation into a Jordanian Disinformation Campaign on Facebook, TikTok, and Twitter', 2021; Victoria Scott, 'Complete Document - Gender Dimensions | Countering Disinformation', 2021, https://counteringdisinformation.org/topics/gender/complete-document-gender-dimensions; François Allard-Huver, 'Fausses informations, réseaux sociaux et élections: Une perspective info-communicationnelle sur le référé fake news', 2021; Allard-Huver; João Canavilhas, Juliana Colussi, and Zita-Bacelar Moura, 'Desinformación En Las Elecciones Presidenciales 2018 En Brasil: Un Análisis de Los Grupos Familiares En WhatsApp', *El Profesional de La Información* 28, no. 5 (14 September 2019), https://doi.org/10.3145/epi.2019.sep.03.

of "spam" content. Such scale is actually a function of the big platforms' own resolve to prioritise growth in the first place, but nevertheless the challenges are still specially highlighted in terms of the sheer volume involved.

For example, YouTube reports that each quarter it removes one billion comments, 9.5 million videos and 2.2 million channels for violating its policies (including a ban on spam).[105] From July to September 2022, the figure was 737,512,355 comments on videos removed, of which two thirds were spam or scams.[106] Google Maps says that, in 2019, it removed more than 75 million policy-violating reviews and four million fake business profiles.[107] TikTok said it removed 102,305,516 videos (about 1% of all video uploads on the platform) over the first quarter of 2021. Twitch reported banning over "13 million disruptive bot accounts". In the first quarter of 2022, Facebook said it deleted 1,6 billion fake accounts.[108] This refers to accounts deemed by the company to be created with malicious intent, or created to represent a business, organisation or non-human entity, and the figure is actually down from 2,2 billion deletions in the first quarter of 2019.[109] The company says it has placed warnings on 200 million distinct pieces of content, based on fact-checking results[110].

TRACKING CHANGES

Much "naïve" misinformation online originates in the efforts of deliberate disinformation actors to spark virality as much as possible. The combined result is a massive proliferation of problematic content online, and this raises the question of benchmarks for assessing how the companies address it. A "misinformation amplification ranking" devised by researchers found that, based mainly on likes and shares of proven misinformation, Twitter and Tiktok have worse scores than Facebook, Instagram and YouTube.[111] Meanwhile, for their own part, the major platforms devise and apply particular metrics to score their own performance in regard to detecting and actioning a range of content that violates their policies, as well as how they performed with appeals against their decisions.[112] However, as Ofcom researchers have pointed out, these figures presented in isolation prompt more questions than answers.[113]

Meta's metrics include how much "violating content" they have found and actioned before users reported it. A further Meta measurement is called "prevalence", which is based on a sampling method to estimate the percentage of total users on the service who may have viewed "violating content". This provides a picture of much harmful content is potentially seen on the platform, rather than how much is posted in total. Such figures can show improvements in company performance over time, although they do not necessarily make sense in the absence of data about changes in the volume and character of posts and engagement rates with this kind of content. More fundamentally, the measure intrinsically produces statistics favourable to the company since the total content views will almost invariably be a much greater number than the estimated figure for any potentially viewed specific item/s of content. The quantitative focus of this metric also diverts attention away from the qualitative relevance of problematic materials that may reach very specific groups of actors bent on violating human rights. Such content can support organising, training, mobilising and fundraising purposes which are all highly likely be at odds with the company's stated goals. To assess this problem would require far more pro-active and nuanced measurement in order to holistically interpret the significance of changing patterns of success, stasis or regress in the company's mitigation record.

As regards Twitter, this platform has operated a metric titled "violative view rate" which records the number of actual views a Tweet considered to violate company policies receives, prior to its removal. Again, however, this is a measurement that privileges corporate achievements in limiting the absolute reach of problematic content to a mass of undifferentiated individuals, while simultaneously leaving aside the challenge of qualitatively assessing the functionalities of such messages at least within communities and groups with interests in disinformation and hate speech.

105 The YouTube Team, 'Managing Harmful Conspiracy Theories on YouTube'.
106 World Federation of Advertisers, 'GARM Aggregated Measurement Report – November 2022'
107 Google Youtube, 'Information Quality & Content Moderation'.
108 Stephen Warwick, 'Facebook removed 1.6 billion fake accounts in just three months', 2022, https://www.imore.com/facebook-removed-16-billion-fake-accounts-just-three-months
109 https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/
110 World Federation of Advertisers, 'GARM Aggregated Measurement Report – Nov 2022'
111 Integrity Institute, 'Misinformation Amplification Analysis and Tracking Dashboard', Integrity Institute, 2022, https://integrityinstitute.org/our-ideas/hear-from-our-fellows/misinformation-amplification-tracking-dashboard.
112 World Federation of Advertisers. GARM Aggregated Measurement Report - November 2022, 2022, https://wfanet.org/leadership/garm/garm-resource-directory-%28weblog-detail-page%29/2022/05/17/GARM-Aggregated-Measurement-Report-November-2022
113 Harling, Anna-Sophie, Declan Henesy, and Eleanor Simmance. 'Transparency Reporting: The UK Regulatory Perspective', Journal of Online Trust and Safety 1, no. 5, 2023, https://tsjournal.org/index.php/jots/article/view/108.

The larger platforms themselves increasingly report cases of manipulation[114], and point fingers at what they call "bad actors" (connoting that they are "good guys"). However, beyond such a Manichean representation, it is also the case that their own systems, policies and practices enable results for the manipulators. For example, while Facebook has a rule that each person should only have one account, researchers have found that a single user can easily make multiple pages – a "loophole" that has allowed political manipulation across 25 countries.[115] Also on Facebook, researchers have found at least 24 of 80 white supremacist groups' Pages were auto-generated by Facebook itself.[116]

More fundamentally, the platforms' tools of audience profiling and targeted advertising have been assessed as facilitating "digital deceit" and "precision propaganda".[117] Like the platforms themselves, external manipulators seek targeting and ongoing engagement with their content. Among the many illustrations of them using the affordances of platforms are the following:

- Organised manipulation on social media has been able to more than double since 2017, with 70 countries using computational propaganda to manipulate public opinion, according to the Oxford Internet Institute. The same research says that politicians in 45 countries have used social media to try to amass fake followers or spread manipulated media to garner voter support.[118]
- In Kenya, researchers uncovered at least 11 different disinformation campaigns, manifested in more than 23 000 tweets.[119]
- In South Africa, "a dangerously orchestrated narrative" whipping up anti-immigrant sentiment, from a "well-oiled propaganda machine" of 80 interconnected Twitter accounts, was able to use the platform's possibilities to reach 50 000 other accounts.[120]
- In India, over 10 000 political volunteers were shown to be servicing 50 000 WhatsApp groups with content produced by a party's social media team, including false messages, inaccurate framing, 'fear speech; and hate speech against minorities.[121]

"At least 81 countries have been using social media to spread disinformation, drawing on 41 private firms selling propaganda as a service. Of these, 79 countries are known to use human-curated accounts, and seven are documented using automated accounts. In 14 countries, there have been hacked, stolen or impersonation accounts. 73% of study countries have employed trolling and harassment in order to suppress participation."

Researchers Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, 2020. [122]

114 Twitter, 'Our Range of Enforcement Options for Violations | Twitter Help', n.d, https://help.twitter.com/en/rules-and-policies/enforcement-options; Meta, 'Threat Report: Combating Influence Operations | Meta', 2021, https://about.fb.com/news/2021/05/influence-operations-threat-report/.

115 Julia Carie Wong, 'Revealed: The Facebook Loophole That Lets World Leaders Deceive and Harass Their Citizens | Facebook | The Guardian', 2021, https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation.

116 Abby Ohlheiser, 'The Most Popular Content on Facebook Belongs in the Garbage | MIT Technology Review', 25 August 2022, https://www.technologyreview.com/2022/08/25/1058662/facebook-widely-viewed-content-spam-memes/.

117 Dipyan Ghosh and Ben Scott, 'Digital Deceit: The Technologies Behing Precision Propaganda on the Internet', 2018; Lise Henric, 'Les fake news, entre outils de propagande et entraves à la liberté de la presse', *Hermès, La Revue* 82, no. 3 (2018): 120–25, https://doi.org/10.3917/herm.082.0120.

118 Samantha Bradshaw et al., *Country Case Studies Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation* (Oxford Internet Institute, 2021).

119 Odanga Madung and Brian Obilo, 'Inside the Shadowy World of Disinformation for Hire in Kenya', 2022, https://counteringdisinformation.org/topics/platforms/complete-document-platforms; Patrick Mutahi and Brian Kimari, 'Fake News and the 2017 Kenyan Elections', *Communicatio* 46, no. 4 (1 October 2020): 31–49, https://doi.org/10.1080/02500167.2020.1723662; Madung Odanga, 'From Dance Ap to Political Mercenary. How Disinformation on TikTok Gaslights Political Tensions in Kenya', 2022; Vittoria Elliot, 'Disinfo and Hate Speech Flood TikTok Ahead of Kenya's Elections | WIRED', 2022, https://www.wired.com/story/kenya-tiktok-election-disinformation-hate-speech/; Patrick Mutahi and Brian Kimari, 'Fake News and the 2017 Kenyan Elections', *Communicatio* 46, no. 4 (1 October 2020): 31–49, https://doi.org/10.1080/02500167.2020.1723662; Billy Mutai, 'Fake Images and Disinformation on Social Networking Sites: Case Study of Kenya's 2017 General Election' (Thesis, University of Nairobi, 2021), http://erepository.uonbi.ac.ke/handle/11295/160578.

120 Jessica Bezuidenhout, '@uLerato_pillay: How the xenophobic network around #PutSouthAfricaFirst was born and then metastasised', 18 August, 2020, https://www.dailymaverick.co.za/article/2020-08-18-ulerato_pillay-how-the-xenophobic-network-around-putsouthafricafirst-was-born-and-then-metastasised/

121 Samriddhi Sakunia, 'Elections in Modi's Home State Have Become a WhatsApp Spam War', Rest of World, 2 December 2022, https://restofworld.org/2022/whatsapp-gujarat-india-elections/.

122 Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation (Computational Propaganda Project at the Oxford Internet Institute, 2021), https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/CyberTroop-Report20-Draft9.pdf

Disinformation-for-hire is increasingly reported, whereby public relations companies sell this service globally, based on their abilities to exploit platform operations.[123]

- In Honduras, a PR firm generated disinformation through Facebook pages and websites making it appear as legitimate news, and a case was exposed in which 96% of 60 000 tweets aiming to discredit opposition figures came from just 41 accounts, of which only 33 were then suspended by Twitter.[124]
- Covert information operations from a Tunisian-based agency seeking to influence African presidential elections involved approximately 3.8 million Facebook accounts following inauthentic pages, with nearly 132,000 joining operation-administrated groups and over 171,000 following the operations' Instagram accounts.[125]

There is also effective exploitation of encrypted communications and closed groups[126] with manipulators using these features to evade formal company restrictions on the use of the services. In response, platforms sometimes deploy meta-data analysis and implement restrictions on behaviours to limit potential content risks.[127] However, the case has been made that Meta could give more support for WhatsApp group admins so that they can, inter alia, moderate the content shared in the groups they create, thereby allowing more user-focused content moderation to combat misinformation and disinformation on the channel.[128]

The situation bodes to become even more challenged in terms of combatting manipulation. Automated efforts continue to 'game' the propensity of algorithms so as to advance virality. "Cheap fakes" are already tolerated on platforms like Facebook (though not YouTube) when used to damage the right to reputation[129]. These will increasingly be supplemented by "deep fakes" that are easily generated by Artificial Intelligence. Meta says it will label or remove this kind of synthetic content except in cases like parody,[130] although it acknowledges detection challenges. Further looking ahead, technologies such as ChatGPT have been flagged as tools that can be cheaply and easily used to generate an ever-increasing deluge of text-based disinformation online.[131]

An overarching concern in regard to manipulation is that many platforms tend to react belatedly. This is due to an absence of systematic prior risk assessments which could enable pre-emptive mitigation measures such as changes to the many vulnerabilities in their systems and also provide at least for stricter enforcement of policies against those promoting content that violates terms of service.

An overarching concern in regard to manipulation is that many platforms tend to react belatedly. This is due to an absence of systematic prior risk assessments which could enable pre-emptive mitigation measures.

123 DiResta et al., 'Mind Farce'; Jonathan Corpus Ong and Katrina Ventura, 'Communication's Jonathan Corpus Ong's "Catch Me If You Can" Podcast Returns for Second Season: UMass Amherst', accessed 18 December 2022, https://www.umass.edu/news/article/communications-jonathan-corpus-ongs-catch-me-if-you-can-podcast-returns-second-season.

124 Leo Schwartz, 'A Prominent PR Firm Is Spreading Disinformation Ahead of Honduras' Elections, New Investigation Reveals', Rest of World, 29 October 2021, https://restofworld.org/2021/political-pr-firm-disinformation-honduras-elections/; Leo Schwartz, 'New Report Reveals Rampant Social Media Manipulation in Honduras' Presidential Elections', Rest of World, 27 May 2021, https://restofworld.org/2021/social-media-honduras-elections/.

125 Andy Carvin, Luiza Bandeira, Graham Brookie, Iain Robertson, Nika Aleksejeva, Alyssa Kann, Kanishk Karan, Ayushman Kaul, Tessa Knight, Jean le Roux, Roman Osadchik, Esteban Ponce de Leon, 'Operation Carthage. How a Tunisian company conducted influence operations in African presidential elections', 2020, Atlantic Council, DFR Lab, https://www.atlanticcouncil.org/wp-content/uploads/2020/06/operation-carthage-002.pdf

126 Tamian Derivry, 'Digital Democracy and Left Party Politics in the 2022 French Presidential Elections', Tech Policy Press, 28 September 2022, https://techpolicy.press/digital-democracy-and-left-party-politics-in-the-2022-french-presidential-elections/; Brandie Nonnecke et al., 'Harass, Mislead, & Polarize: An Analysis of Twitter Political Bots' Tactics in Targeting the Immigration Debate before the 2018 U.S. Midterm Election', *Journal of Information Technology & Politics* 19, no. 4 (2 October 2022): 423–34, https://doi.org/10.1080/19331681.2021.2004287; Sanjana Hattotuwa, Yudhanjaya Wijeratne, and Raymond Serrato, 'Weaponising 280 Characters', 2018.

127 Seny Kamara et al., 'Outside Looking in. Approaches in End-to-End Encrypted Systems', n.d.; Tech Agaist Terrorism, 'Analysis: ISIS Use of Smaller Platforms and the DWeb to Share Terrorist Content', 29 April 2019, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/, https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/.

128 Michael Rain, 'Give Group Admins Tools to Fight Disinformation In Immigrant Diaspora WhatsApp Groups' 2022. https://techpolicy.press/give-group-admins-tools-to-fight-disinformation-in-immigrant-diaspora-whatsapp-groups/

129 Stephanie MacLellan, 'Moderating content in the age of disinformation', 2019, https://www.cigionline.org/articles/moderating-content-age-disinformation/

130 Monika Bickert, 'Enforcing against manipulated media', 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

131 Olivia Solon, 'ChatGPT — Eloquent Robot or Misinformation Machine?: QuickTake', 2023, https://www.washingtonpost.com/business/chatgpt—eloquent-robot-or-misinformation-machine-quicktake/2023/01/12/05da34a6-92c8-11ed-90f8-53661ac5d9b9_story.html ; Steve Jones, 'GPT Chat and the weaponization of disinformation', 2022, https://blog.metamirror.io/gpt-chat-and-the-weaponization-of-disinformation-4701d11c61a0; Cade Metz, 'The New Chatbots Could Change the World. Can You Trust Them?', 2022, https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html; Cade Metz and Scott Blumenthal, 'How A.I. Could Be Weaponized to Spread Disinformation', 2019, https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html?action=click&module=RelatedLinks&pgtype=Article

COMPANY SPENDING PRIORITIES

For many platforms, the goal is untrammelled pursuit of growth and profit[132], with policy and practice being driven by business concerns, even if at the expense of safety. This characteristic has even been noted by Meta's own "oversight board" regarding the company's cross-check policy[133]. What this particular prioritisation leads to is inadequate attention to potential risks and a corresponding underinvestment in at least mitigating the effects of the core business model. [134] Scant resources are allocated to building moderation capacity in various languages in markets regarded as less lucrative to the company concerned (see Part 3 in this series). Percentage-wise, it appears that little is likely spent in exploring alternative ways to curate and moderate content[135], or in experimenting with different methods to generate revenues, [136] while costly acquisitions of potential competitors or contributors to the ongoing basic business model are frequent. Amongst other corporate behaviours, the treating of data holdings as exclusive private assets serves to disadvantage independent safety-tech developers who argue that "While institutions holding data are responsible for safeguarding and protecting from misuse, this builds walls around datasets and makes it challenging for Safety Tech to test and further develop". [137]

Many companies argue that they do spend sizeable sums on their moderation efforts, and also allocate monies (although the figures are typically not disclosed) to fact-checking and to media and information literacy efforts. At the same time, these very entities are also shown to be dispensing what are likely to be very sizeable amounts[138] on advocating for "tech solutionism"[139] and on other policy lobbying[140] and self-serving sponsorships.[141] Research has shown that Google funded 330 research papers published between 2005 and 2017 on public policy matters of interest to the company.[142] The company has also been accused of running a stealth influence campaign.[143] Other commercial platforms sponsor third parties as part of influence efforts.[144] Facebook is reported to run its own influence campaigns.[145] Some observers see platforms claiming enthusiasm for their operations being regulated by law, while exploiting the current lack

Percentage-wise, it appears that little is likely spent in exploring alternative ways to curate and moderate content, or in experimenting with different methods to generate revenues.

132 Aviv Ovadya, 'Bridging-Based Ranking. How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy', Belfer Center for Science and International Affairs, 2022, https://www.belfercenter.org/publication/bridging-based-ranking; Nathalie Maréchal and Ellery Roberts Biddle, 'It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge', 2020.

133 Oversight Board, 'Policy Advisory Opinion on Meta's Cross-Check Program', n.d.

134 Sheera Frenkel and Cecilia kang, 'Amazon.Com: An Ugly Truth: Inside Facebook's Battle for Domination', 2021, https://www.amazon.com/Ugly-Truth-Inside-Facebooks-Domination/dp/0062960679; Nathalie Maréchal, 'We Can't Govern the Internet without Governing Online Advertising. Here's How to Do It. - Ranking Digital Rights', Ranking Digital Rights, 2022, https://rankingdigitalrights.org/mini-report/we-must-govern-online-ads/.

135 Daphne Keller, 'Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users', The University of Chicago Law Review Online, 28 June 2022, https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/; Robyn Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches', 2018.

136 Christian Djeffal, Christina Hitrovaf, and Eduardo Magrani, 'Recommender Systems and Autonomy: A Role for Regulation of Design, Rights, and Transparency', 2021.

137 Online Safety Tech Industry Association, 'The International State of Safety Tech An Analysis of Global Market Trends', n.d., https://view.publitas.com/public-1/international-state-of-safety-tech/page/1

138 Corporate Europe Observatory. Big Tech's last minute attempt to tame EU tech rules, https://corporateeurope.org/en/2022/04/big-techs-last-minute-attempt-tame-eu-tech-rules; Mie Oehlenschlager, 'Big Tech's Soft Power To Be Investigated in Brussels', 2023, https://dataethics.eu/big-techs-soft-power-to-be-investigated-in-brussels/; Clothilde Goujard, 'Big Tech accused of shady lobbying in EU Parliament', 2022, https://www.politico.eu/article/big-tech-companies-face-potential-eu-lobbying-ban/;

139 Pawel Popiel, 'Digital Platforms as Policy Actors', in *Digital Platform Regulation: Global Perspectives on Internet Governance*, ed. Terry Flew and Fiona R. Martin, Palgrave Global Media Policy and Business (Cham: Springer International Publishing, 2022), 131–50, https://doi.org/10.1007/978-3-030-95220-4_7.

140 Vicky Cann and Nina Katzemich, 'Big Tech Now Edges out Big Energy in EU Lobbying | Corporate Europe Observatory', 2022, https://corporateeurope.org/en/2022/09/big-tech-now-edges-out-big-energy-eu-lobbying; Pietro Lombardi, 'Big Tech Boosts Lobbying Spending in Brussels', *POLITICO* (blog), 22 March 2022, https://www.politico.eu/article/big-tech-boosts-lobbying-spending-in-brussels/; Corporate Europe Observatory, 'Big Tech Now Edges out Big Energy in EU Lobbying | Corporate Europe Observatory', *Corporate Europe Observatory* (blog), 2022, https://corporateeurope.org/en/2022/09/big-tech-now-edges-out-big-energy-eu-lobbying.

141 Nicolas Kayser-Bril, 'How Big Tech Charms and Bullies European Politicians, Journalists and Academics', AlgorithmWatch, 2021, https://algorithmwatch.org/en/big-tech-lobby-influence-europe/.

142 Tech Transparency Project, 'Google Academics Inc.', Tech Transparency Project, 11 July 2017, https://www.techtransparencyproject.org/articles/google-acc.

143 Tech Transparency Project, 'Google Targets the Left with Stealthy Influence Campaign', Tech Transparency Project, 28 May 2022, https://www.techtransparencyproject.org/articles/google-targets-left-stealthy-influence-campaign.

144 Tech Transparency Project, 'Find Out Which Groups Get Big Tech Funding', Tech Transparency Project, 10 August 2021, https://www.techtransparencyproject.org/articles/find-out-which-groups-get-big-tech-funding.

145 Tech Transparency Project, 'Funding the Fight Against Antitrust: How Facebook's Antiregulatory Attack Dog Spends Its Millions | Tech Transparency Project', 2022, https://www.techtransparencyproject.org/articles/funding-fight-against-antitrust-how-facebooks-antiregulatory-attack-dog-spends-its-millions.

of constraints and opposing actual steps.[146] In this context, mandatory disclosures of corporate lobbying are limited or non-existent in many jurisdictions.

Some platforms have been making changes such as phasing out "3rd party tracking cookies" which will reduce their power to harvest user-data from across the wider online ecosystem. This will in turn reduce some of their ability to target content and adverts to 'users', and as such is seen by stock markets as reducing platform revenue prospects. However, such changes are only being introduced in the wake of the EU's General Data Protection Regulation, with no evidence that this will change platform use of data that is directly and indirectly collected (such as being bought-in), in order to drive a system which favours growth and financial returns even in the face of an imbalance in terms of spending to better protect human rights.

STAKEHOLDER KNOWLEDGE DEFICITS

There are sub-optimum levels of governmental knowledge in many countries about the complexities of the platforms' operations. Citizens' levels of media and information literacy, especially in the digital realm, are also low.[147] While media and information literacy can be a part antidote to potentially harmful content, there is relatively little available evidence of the extent or impact of activity. Under current governance arrangements, it can be assessed that neither governments nor platforms appear to be systematically working to achieve society-wide impact to promote such competencies.[148] Studies of digital readiness show there is much to be done, not least with newer users and elderly people. A Media Literacy Index published in 2022 found potential vulnerability to misinformation of 41 societies in Europe.[149]

Several companies have activities to educate and empower users in the safe use of their services, but again it is hard to assess these efforts from the outside in the absence of data, although the impact may be known to the platforms themselves. There is also concern that companies should not shift the burden of responsibility onto individuals to build personal resilience and skills to navigate the totality of online expression (including plausible but actually synthetic content), while the platforms themselves do too little 'upstream' to limit the flood of potentially harmful content before it reaches vulnerable people.[150] At the same time, there is little public literacy about the possible negative implications for freedom of expression by private gatekeepers, such as in regard to blocking "false positives" through automated filtering – with associated curbs on legitimate content (like journalism) in terms of upload or reach.

Aggravating low knowledge levels about how platforms operate is the lack of corporate transparency amongst these enterprises. This opacity can also thwart the development and implementation of effective regulation frameworks. Meta's own 'oversight board' has itself complained that it continues to lack data from the company to verify progress on how Meta is implementing board recommendations, adding that "there is a lack of transparency around how Meta's automated systems work and how they affect the content users see".[151]

---

146 Rebekah Tromble, 'Facebook, NYU, and the "Risks" of Public Interest Research', Tech Policy Press, 10 August 2021, https://techpolicy.press/facebook-nyu-and-the-risks-of-public-interest-research/; Terry Flew and Fiona R. Martin, eds., *Digital Platform Regulation: Global Perspectives on Internet Governance*, Palgrave Global Media Policy and Business (Cham: Springer International Publishing, 2022), https://doi.org/10.1007/978-3-030-95220-4.

147 Michael Pal, 'Social Media and Democracy: Challenges for Election Law and Administration in Canada', *Election Law Journal: Rules, Politics, and Policy* 19, no. 2 (1 June 2020): 200–213, https://doi.org/10.1089/elj.2019.0557; Jamie Hitchen et al., 'Whatsapp and Nigeria's 2019 Elections : Mobilising the People, Protecting the Vote', Africa Portal (Centre for Democracy and Development (CDD), 1 July 2019), https://www.africaportal.org/publications/whatsapp-and-nigerias-2019-elections-mobilising-people-protecting-vote/; Cunliffe-Jones et al., *Misinformation Policy in Sub-Saharan Africa*.{\\i{}Election Law Journal: Rules, Politics, and Policy} 19, no. 2 (1 June 2020

148 Lukman Mahami Adams, 'Understanding the Concept of Political-Fact Checking in an Election Year': (Ghana, Ghana Institute of Journalism, 2021), https://repository.gij.edu.gh/bitstream/handle/gijdr/269/Understanding%20The%20Concept%20f%20Political%20Fact-Checking%20In%20An%20Election%20Year%20A%20Comparative%20Analysis%20f%20Political%20Fact-Checking%20By%20Dubawa%20Ghana%20And%20Ghana%20Fact%20During%20Ghana%E2%80%99s%202020%20General%20Election.pdf?sequence=1&isAllowed=y; Cunliffe-Jones et al., *Misinformation Policy in Sub-Saharan Africa*.Ghana Institute of Journalism, 2021

149 Marin Lessenski, 'How It Started, How It Is Going: Media Literacy Index 2022' (Open Society Institute, n.d.), https://osis.bg/wp-content/uploads/2022/10/HowItStarted_MediaLiteracyIndex2022_ENG_.pdf.

150 Abhishek, 'Overlooking the Political Economy in the Research on Propaganda'.scholars studying propaganda have focused on its psychological and behavioral im-pacts on audiences. This tradition has roots in the unique historical trajectory of the United States through the 20th century. This article argues that this tradition is quite inadequate to tackle prop-aganda-related issues in the Global South, where a deep understanding of the political economy of propaganda and misinformation is urgently needed.","container-title":"Harvard Kennedy School Misinformation Review","DOI":"10.37016/mr-2020-61","journalAbbreviation":"HKS Misinfo Review","source":"DOI.org (Crossref

151 Oversight Board, 'Oversight Board Publishes Transparency Report for Second Quarter of 2022 and Gains Ability to Apply Warning Screens | Oversight Board' (Oversight Board, 2022), https://www.oversightboard.com/news/784035775991380-oversight-board-publishes-transparency-report-for-second-quarter-of-2022-and-gains-ability-to-apply-warning-screens/; Oversight Board, 'Oversight Board Announces Seven Strategic Priorities | Oversight Board', 2022, https://www.oversightboard.com/news/543066014298093-oversight-board-announces-seven-strategic-priorities/.

4 Recommendations

- Guidance emphasising human rights as the appropriate vantage point for assessing problems and regulatory solutions can help to entrench these international standards as foundational for any regulatory regimes

- Any law-based resort to a regulatory 'cure' for the platform ills must be structured to avoid worsening the 'disease', given that state actors are frequently implicated in contributing to online content that threatens human rights and may overstep their roles regarding the control of online content.

- Guidance can advise that regulating only the largest platforms is insufficient, even if at the same time regulatory regimes should be nuanced in terms of platform size and role, and encryption-enabling privacy should not be compromised in regard to messaging services

- Guidance can be given about the need to take stock of the role of 'attention economics' business models, automated advertising and supporting a pluralism of platforms that includes a range of non-profit business models.

- Guidance can encompass both curation and moderation and require platforms to do what it takes to better in both areas to achieve results in addressing misinformation and hate, with comprehensive metrics agreed on what constitutes appropriate levels of achievement.

- Focus can be put on the need for platforms to define and identify manipulative influence operations which traffic in lies and hatred, while also requiring platforms to respect the privacy of users (including the right to anonymity) and allowing for legitimate freedom of expression as core to information as a public good.

- Guidance can highlight the need for platforms to increase their "clean up" spending such as through resourcing more fact-checking and providing support to media and information literacy initiatives, with the development of appropriate metrics as benchmarks for what is appropriate in the countries of company operation.

- Recognising the importance of independent monitoring of the problems and attempted solutions, there should be mandated transparency as a key component of guidance. This could propose practical enforcement through documentation, data, code and company research requirements[152]. This element would enable external scrutiny, and even independent audit, of platform claims, and could also help to stimulate research and reform.

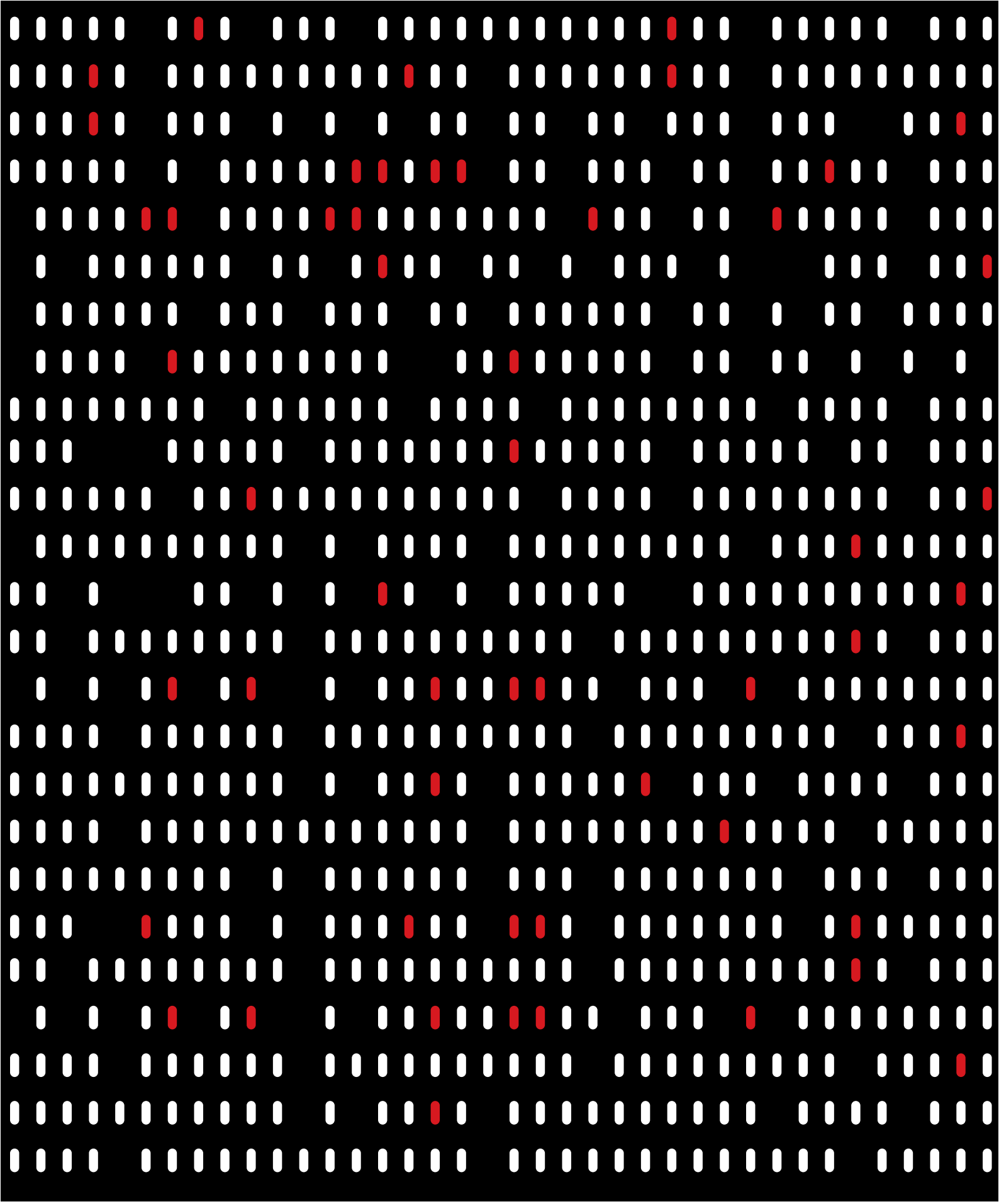
152 Priyanjana Bengani, 'A Menu of Recommender Transparency Options', Tech Policy Press, 8 August 2022, https://techpolicy.press/a-menu-of-recommender-transparency-options/.

5 Call for input

Would you like to comment on this working document? We'd especially like to hear your views on:

- Are there inaccuracies or omissions?
- Are there ways to conceptualise a distinction between content that potentially harms human rights, and content that intrinsically breaches such rights, and with what regulatory implications?
- How can the issue of transnational platforms versus national jurisdictions be tackled, especially as regards how small and developing countries can configure appropriate regulatory solutions (keeping in mind that this issue is also given more attention in Part 3 of this series)?
- How could proportionality criteria be elaborated in terms of regulating different-sized platforms, in the light of a basic threshold which applies even to small platforms? (Again, (keeping in mind that this issue is also given more attention in Part 3 of this series).
- What process can be proposed to develop better metrics for assessing problems and companies' attempts to address them?
- Is there scope for mitigating the attention-economics business model and data-driven targeted advertising or if regulatory arrangements should aim at more fundamental change? How?
- Is there scope for reforming the current configurations of automated ad-tech, and if so how?
- Is there opportunity to recommend that content prioritisation of short entertainment video style be complemented by provision of feeds that encourage people to choose alternatives that optimise for quality informative and educational content?
- How could regulation foster alternatives to the model of commercial companies focused on growth and financial returns?

Comments can be sent to e-mail: internetconference@unesco.org with the subject line: *Response to draft background paper or via this form here:* https://forms.gle/iHeddmLwWEMyXXUo7



This research was supported by UNESCO