PART 2

DIGITAL PLATFORM GOVERNANCE AND THE CHALLENGES FOR TRUST AND SAFETY

RESEARCH ICT AFRICA

PLATFORMS' OWN POLICIES AND PRACTICES: WHAT PROBLEMS NEED CHANGING?

KEY TRENDS UNCOVERED

- Platform policies lack clarity about the relationship between them, and also about how policies should be applied at global and local levels.
- How platforms understand and identify harms is insufficiently mapped to human rights standards, and there is a gap in how policy elements should deal with different rights or with business models when there are tensions.
- Policies are not always transparent and do not provide sufficiently for risk assessment.
- Implementation and enforcement by platforms have serious shortfalls, while attempts to improve outcomes by automating moderation have their limits.
- Inequalities in policy and practice abound in relation to different categories of people, countries and languages.
- Of value in addressing these problems could be the development of guidance for the governance and regulation of frameworks that sets out suggested standards and parameters for platform policies and related operations.

This is a draft background paper developed with the support of UNESCO by Research ICT Africa, a digital policy, regulation and governance 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 2 of a three-part series. The whole 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 these pages 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.

One of the major global governance challenges today is mitigating the risks associated with online disinformation, misinformation and hate speech. Platforms' policies are supposed to perform at the heart of how these entities deal with such content-related issues. The policies ought to encompass problems resulting from foundational curational design of their services. As a service commitment, policies should constitute a legal promise to consumers, and they should provide a safety net with clear and predictable rules. They should be effectively implemented, assessed and renovated. While recognising diverse performance across different platform companies, this Part 2 of the series examines common key problems and some successes, and points out where action is needed.

1 Core gap in platform policies: international human rights standards

Digital platforms' content policies tend to be global and generic, with few customisations according to national or regional requirements. Mark Zuckerberg of Meta has said that his policy team is based across six countries, adding that this is "to reflect the different cultural norms of our community".[1]

But without further detail, it is not possible to know how this assertion modifies a "one size fits all" approach. The company says that its "Community Standards apply to everyone all around the world, and to all types of content".[2] For its part, YouTube says its Community Guidelines "apply to everyone, and to all types of content on YouTube - such as videos, comments, links, and thumbnails".[3] It elaborates: "YouTube's Community Guidelines are enforced consistently across the globe, regardless of where the content is uploaded. When content is removed for violating our guidelines, it is removed globally."

At the same time, platforms also tailor the enforcement of their policies to accommodate local laws in some cases, as well as operate different tiers based on age and/or political status, amongst other criteria. The issue is whether such divergences of the local from the global remain within a company's wider policy claims and parameters (as would be assessed, for example, when investigating compliance with international human rights standards in different jurisdictions around the world). In the absence of companies having clear elaboration of how their policies operate at global and local levels, it is to be expected that there will be many international inconsistencies in application – some of which are cited below.

Another area lacking elaboration is that companies refer to a number of specific types of "harms" as if these were self-evident. But this approach neglects to link these to the panoply of distinctive human rights (such as expression, dignity and personal autonomy), and it is thus unable to conceptualise safety as the condition when individuals and society are protected from potential harms to these rights by platform content and communications. The same issue characterises some legal regulations which list a range of concrete harms without wider human rights reference[4], meaning, inevitably, that new risks and unanticipated threats to safety are likely to escape the list of current concerns that are concretised in very specific legal formulations but lack a more general beacon.

The kinds of relevant standards that should apply to content management include treaties like the International Covenant on Civil and Political Rights (ICCPR) and the International Convention on the Elimination of All Forms of Racial Discrimination. In addition, there is relevant guidance from treaty bodies such as the Human Rights Committee's General Comments 34 and 25. Also pertinent are the reports of the UN special rapporteur on freedom of expression and opinion, and those of the UN special rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance. The UN's Rabat Plan of Action against hate speech, developed by the Office of the High Commissioner for Human Rights, offers an appropriate way to apply related policies by elaborating how and when there is a case

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

2 Meta, 'Inauthentic Behaviour | Transparency Centre', n.d., https://transparency.fb.com/en-gb/policies/community-standards/inauthentic-behavior/.

3 Youtube, 'YouTube Community Guidelines & Policies - How YouTube Works', YouTube Community Guidelines & Policies - How YouTube Works, n.d., https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.

4 Ireland, 'Online Safety and Media Regulation Act 2022', https://data.oireachtas.ie/ie/oireachtas/act/2022/41/eng/enacted/a4122.pdf; Australia, 'Basic Online Safety Expectations esafety.gov.au Regulatory Guidance', 2022, https://www.esafety.gov.au/sites/default/files/2022-07/Basic%20Online%20Safety%20Expectations%20regulatory%20guidance.pdf

for restrictions.[5] While many platform's policies make formal reference to human rights,[6] most share the feature that they are connected inadequately to these elaborated international standards on freedom of expression.[7] One exception is Meta's "oversight board" which references more than 15 international standards in explaining its assessments, and which has also noted Facebook referencing the Rabat Plan in one content moderation decision[8]. In the case of Alphabet (owner of Google and YouTube amongst other companies), the company commits to human rights and its board has amended the charter of its Audit and Compliance Committee to include human rights.[9] However, this appears to have not enjoyed material follow up in subsequent company statements. The same documents do show that in 2021 the Alphabet board recommended against a shareholder proposal to nominate a human rights expert to its ranks, as it did also in regard to a proposal for a third-party review of the company's whistle-blower policies.[10]

Significantly, platforms' policies at large also do not explicitly encompass human rights in regard to advertising and other paid-for content,[11] and in Meta's case they explicitly entail a broadly hands-off approach to disinformation in political adverts even though this kind of content can subvert the right to political participation.

Many large platform companies pledge support to the UN Guiding Principles on Business and Human Rights which call for conducting human rights due diligence and disclosure thereof. However, in the absence of legal requirements, it is left for companies to self-decide if and when to undertake such exercises.

Platforms in the Global Network Initiative[12] sign up to principles on freedom of expression and privacy. The focus of this association is for its corporate members to "work to protect the freedom of expression rights of users when confronted with government demands, laws and regulations to suppress freedom of expression, remove content or otherwise limit access to communications, ideas and information in a manner inconsistent with internationally recognised laws and standards". Such an approach means that risks are construed primarily in terms of governmental action, thereby underplaying risks from other external actors, internal business models, and insufficient spend on capacity to moderate content in the languages of markets entered. These features qualify the impact that such corporate commitments to human rights could have.

> With a few exceptions, companies seldom reference international human rights standards as a basis for their policies. Instead, they draw on their numerous separate policies to craft responses to conflict-based challenges. This fragmented approach fails to provide much-needed coherence and predictability to platform practice and has the potential to undermine company compliance with international human rights law and international humanitarian law."
> - UN Special Rapporteur on Freedom of Expression and Opinion, Irene Khan, 2022.[13]

5 OHCHR, 'OHCHR | Annual Thematic Reports', OHCHR, 2021, https://www.ohchr.org/en/special-procedures/sr-freedom-of-opinion-and-expression/annual-thematic-reports; OHCHR, 'OHCHR | General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment', OHCHR, accessed 20 December 2022, https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation; OHCHR, 'Guiding Principles on Business and Human Rights' (OHCHR, n.d.), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

6 Ranking Digital Rights, 'The 2022 Ranking Digital Rights Big Tech Scorecard', Ranking Digital Rights, 2022, https://rankingdigitalrights.org/bts22/indicators/P1b.

7 OHCHR, 'Guiding Principles on Business and Human Rights. Implementing the United Nations 'Protect, Respect and Remedy; Framework', 2011.

8 OHCHR, 'OHCHR | The Rabat Plan of Action', OHCHR, 2012, https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action.

9 Google, 'About Human Rights at Google - Google', About Google, n.d., https://about.google/human-rights/.

10 Alphabet, 'Proxy Statement Pursuant to Section 14(A) of the Securities Exchange Act of 1934 (Amendment No.)', accessed 19 December 2022, https://www.sec.gov/Archives/edgar/data/1652044/000130817921000256/lgoog2021_def14a.htm#lgooga048.

11 Justin Hendrix, 'Fake Net Neutrality Comment Campaign a Harbinger of Automated Disinformation to Come', Tech Policy Press, 6 May 2021, https://techpolicy.press/fake-net-neutrality-comment-campaign-a-harbinger-of-automated-disinformation-to-come/.

12 GNI, 'The GNI Principles', Global Network Initiative, n.d., https://globalnetworkinitiative.org/gni-principles/.

13 Irene Khan, 'Disinformation and freedom of opinion and expression during armed conflicts', 2022, https://documents-dds-ny.un.org/doc/UNDOC/GEN/N22/459/30/PDF/N2245930.pdf?OpenElement

The human rights lacunae also mean that many policy sanctions do not optimally provide for the criteria of necessity and proportionality. This shortfall accounts for why the Council of Europe expressly recommends to platforms that "any restriction of content should be carried out using the least restrictive technical means and should be limited in scope and duration to what is strictly necessary to avoid the collateral restriction or removal of legal content."[14] Another consequence of not fully integrating the human rights perspective is that policies also often fall short in terms of elaboration about their own responsiveness and redress commitments.

A related area where policy would benefit from human rights elaboration is how a platform entity will decide in regard to using one policy rather than another. An example is a policy on hate speech vs on incitement to violence, which requires that the company itself is clear about the distinctiveness of hatred against groups based on features like nationality or ethnicity, in comparison to the issue of incitement to violence (such may be in defence of the right to security and national sovereignty, and not predicated on hate speech).[15] Without such clarity, one policy may be inappropriately deployed to do duty in the place of another more suited to the purpose.

The other side of this coin in having clarity about the inter-relationships between policies from a human rights point of view is shown by a case in 2018. Here, it was reported that Facebook had allowed praise of white nationalism and separatism "as an ideology" while expressions of white supremacy were to be prohibited, in this way implying that these elements were meaningfully separable.[16]

It was reported that Facebook had allowed praise of white nationalism and separatism "as an ideology" while expressions of white supremacy were to be prohibited, in this way implying that these elements were meaningfully separable.

Also weak is how company policies differentiate between various rights that can be harmed by hate speech, incitement to violence and certain cases of misinformation (all of which expressions have different statuses under international law under the ICCPR). The rights to safety, equality, health, political participation, etc. are at stake in how companies identify potentially harmful content and which policies to apply to it.

The literature relevant to platform regulation notes that platforms differ a lot in how they define, recognise and treat expressions such as hate speech, misinformation and disinformation.[17] While pluralism amongst platforms is to be valued, the high degree of heterogeneity also suggests that there is an absence of wider and more fundamental standards that could serve as an overall reference point. It follows that such a gap could be usefully filled by regulatory guidance principles which highlight the beacon of international human rights standards. Such guidance might also have resonance with the governance of other players in the "tech stack" which, as noted, can shape particular content and communicators through their processes and policies.

Policies also have little that recognises how to deal with content items that individually may not harm such rights, but which together can amplify, spread and fuse with other elements to constitute a narrative which breaks through a threshold of danger to human rights - which would then violate international standards as well as likely have reached a level of illegality in many countries. An example is platforms allowing conspiracy theories about the Brazilian elections to go viral in the three months before the January 8 storming of key state institutions in the country, with related content being able to gain more than 22 million views on YouTube and 2,3 million Facebook interactions.[18]

This policy challenge can be partly understood by assessing YouTube's stance on what the company calls "borderline videos" [19] This is content which the company assesses is not egregious enough to be removed, but which can attract less harsh moderation like disablement of sharing, commenting, liking and recommending, or warning messages or demonetisation. It may be that the company has internal criteria

14 Council of Europe, 'Recommendation CM/Rec (2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries'.

15 Oversight Board, 'Oversight Board Submission to the Special Rapporteur on Freedom of Opinion and Expression: Challenges in Times of Conflicts and Disturbances', 2022; Ryan Mac, Mike Isaac, and Sheera Frenkel, 'How War in Ukraine Roiled Facebook and Instagram', The New York Times, 30 March 2022, sec. Technology, https://www.nytimes.com/2022/03/30/technology/ukraine-russia-facebook-instagram.html; Will Oremus, 'Analysis | Social Media Wasn't Ready for This War. It Needs a Plan for the next One.', Washington Post, 26 March 2022, https://www.washingtonpost.com/technology/2022/03/25/social-media-ukraine-rules-war-policy/.

16 Paul Barrett M, 'Who Moderates the Social Media Giants? A Call to End Outsourcing' (Center for Business and Human Rights, 2020).

17 Ofcom, 'The Buffalo Attack: Implications for Online Safety', Ofcom, 12 October 2022, https://www.ofcom.org.uk/research-and-data/online-research/the-buffalo-attack-implications-for-online-safety.

18 SumOfUs, 'URGENT REPORT. How Meta and Google enabled and profited from the terrorist attacks in Brazil's capital'

19 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/.

as to how it deems such content to be in this kind of grey zone between being removed or subjected to other treatments. However, if they exist, these criteria do not appear to be public. Nor is it clearly disclosed on what basis YouTube will make use of different moderation treatments. Without such transparency, stakeholders are left to trust that YouTube is simply to be relied on to be consistent and appropriately nuanced in terms of how it identifies and processes content, even when this is escalating towards real danger to human rights.

This all shows that which policy applies when, and to what extent, is a matter that needs work by even the biggest players. But even more in need of attention is when moderation decisions are taken by platforms in isolation of clear policy provisions, therefore counting as arbitrary.[20] Further, where policy implementation relies on human judgement, there is evidence of moderators being insufficiently trained in how to apply policy to specific instances, which results in a number of content takedowns also being assessed as arbitrary, rather than rule-governed as required by human rights standards.[21]

20 19 Oversight Board, 'Case Decision 2021-001-FB-FBR' (Oversight Board, 2021), https://www.oversightboard.com/sr/decision/2021/001/pdf-english.

21 Access Now, '29 Recommendations on Content Governance' (Access Now, 2020).

2 Additional problems with platforms' policies

There is evidence of how platforms' content policies and practices, similarly to the core algorithms, are profoundly shaped by prevalent business models. For example, Facebook and Instagram's "cross-check" system privileged powerful actors and allowed business interests to influence content decisions, leading the company's "oversight board" director to comment that the mechanism "prioritises users of commercial value to Meta and as structured does not meet Meta's human rights responsibilities and company values".[22] Policies that are shaped in this way, and in secrecy, work against the human right to equality.

The large platforms like to draw attention to what they are doing in regard to their moderation efforts (human and technical) to enforce their policies. However, they are generally less forthcoming about problems linked to curation which, as discussed in Part 1, uses algorithms to determine content prioritisation, deprioritisation, discoverability and recommendations.[23] This is borne out by Ranking Digital Rights which revealed massive shortfalls in the performance of 14 leading platforms regarding the impact of their algorithms on human rights.[24]

In this way, platforms elide tensions that arise between curational design and policy objectives. Thus, a major blind spot in regard to their policies is how the company should moderate content violates their terms of service, when business-driven algorithms promote the self-same material in various ways[25]. In these cases, the commercial logic works in a different direction to formal policies covering the permitted content and allowed users.[26]

An example is Facebook which says it seeks to reduce content that is "spammy, sensational or misleading,"[27] but it is exactly such content that gains algorithmic traction. For example, in 2018 the platform introduced an algorithmic change called "Meaningful Social Interactions" optimising for network closeness (friends and family) and for engagement. However, the latter dynamic trumped the former, meaning that misinformation and toxic content came to prevail. When engagement took the form of resharing, it spreads exponentially. Though Facebook staffers then proposed ways to limit virality in general, the company decided to shelve these for what came to be known as emergency "break glass" rare occasions, because of commercial concerns of limiting user engagement.[28] Another example of this tension is where Facebook encourages engagement by auto-generating pages for groups, seemingly without sufficient discrimination, which benefits extremist and terrorist entities who produce materials that violate the platform's own content policies.[29]

Another issue for platforms is how they observe the letter and spirit of their policies. While Facebook deplatformed former US President Donald Trump for a period for very serious policy violations, it allowed his political action committee to continue to advertise with disinformation about the 2020 presidential

22 Katie Paul, 'Meta Oversight Board Calls for Overhaul of Controversial "cross-Check" System for VIPs', Reuters, 6 December 2022, sec. Technology, https://www.reuters.com/technology/meta-oversight-board-calls-overhaul-controversial-cross-check-system-vips-2022-12-06/.
23 Council of Europe, 'Prioritisation Uncovered. The Discoverability of Public Interest Content Online', 2020.
24 Ranking Digital Rights, 'The 2022 Ranking Digital Rights Big Tech Scorecard', 2022.
25 Luke Thornburn, 'What Will "Amplification" Mean in Court?', Tech Policy Press, 19 May 2022, https://techpolicy.press/what-will-amplification-mean-in-court/.
26 Accountable Tech, 'Ban Surveillance Advertising', Accountable Tech, n.d., https://accountabletech.org/campaign/ban-surveillance-advertising/.
27 Henry Silverman and Lin Huang, 'Fighting Engagement Bait on Facebook', Meta (blog), 18 December 2017, https://about.fb.com/news/2017/12/news-feed-fyi-fighting-engagement-bait-on-facebook/.
28 The Journal, 'The Facebook Files, Part 1: The Whitelist - The Journal. - WSJ Podcasts', Wall Street Journal, 2021, https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-1-the-whitelist/aa216713-15af-474e-9fd4-5070ccaa774c.
29 Tech Transparency Project, 'Facebook Profits from White Supremacist Groups'.

election and fundraising for him.[30] This practice generates revenue for the company, but it puts in question the sincerity of their moderation practices. Likewise, leadership claims about moderation at Twitter, that are challenged over their veracity, can undermine public trust in the company's policy enforcement.[31]

Thus, besides prima facie policy, there are many factors, some even contradictory, at work with impact on what content exists, is found, ordered and shared on platforms. Even in the case of encrypted social messaging services, how companies interpret their business interests has impact on content deemed to violate their terms of service. For example, curbs may be set on the frequency and reach of forwarding potential (such as on Whatsapp[32]) - or not set (as generally in the case of Telegram[33]), affecting their consumer propositions in the marketplace and their affordances for content that is contrary to human rights.

BLACK BOXES

Limits within policies and practices mean there is much opacity and generally a lack of granular transparency availed by most platforms. There is selective access to data by a very small number of hand-picked researchers. Criteria about who is a bona fide academic (or NGO researcher) are being discussed in joint fora, but remain to be taken up by most of the platforms concerned.[34] There are also highly uneven practices in terms of keeping open and comprehensive archives of paid political expression.[35] These features have the effect of reinforcing a lack of accountability because they render the platforms as black boxes whose magical algorithmic operations defy full understanding or control. As researchers have argued, even automated content processing (whether by curational or moderational logic) is based upon discretionary prior decisions that are not purely technical. These decisions cascade across the stages of assembling training data, training classifiers, evaluating performance and translating all this into real-world action.[36] It has been said that platforms' particular interests are expressed behind an algorithmic veil.[37]

Additional policy problems that have surfaced are cases of companies operating undisclosed policies, including "shadow banning" [38] and Meta's hidden (but later leaked) system of "cross-checking" that exempts certain people from the processes and policy adjudication applicable to ordinary users.[39] The Meta "oversight board" has noted the company has not been fully forthcoming on this policy, adding: "Meta does not inform users that they are on cross-check lists and does not publicly share its procedures for creating and auditing these lists."[40] The "oversight board" further advised: "The same rules should apply to all users of the platform", although also noting that "what is important is the degree of influence that a user has

30 Michael Scherer and Josh Dawsey, 'Trump looks to 2024, commanding a fundraising juggernaut, as he skirts social media bans', 2021. https://www.washingtonpost.com/politics/trump-fundraising/2021/10/29/5b5a2e64-31b1-11ec-a1e5-07223c50280a_story.html

31 Center for Countering Digital Hate, 'Fact check: Musk's claim about a fall in hate speech doesn't stand up to scrutiny', 2022, https://edition.cnn.com/2022/12/20/tech/zuckerberg-cambridge-analytica/index.html; Brian Fung, 'Zuckerberg weighed naming Cambridge Analytica as a concern in 2017, months before data leak was revealed', 2022, https://edition.cnn.com/2022/12/20/tech/zuckerberg-cambridge-analytica/index.html

32 Vera Zakem, Kip Wainscott, and Daniel Arnaudo, 'Platform Specific Engagement for Information Integrity', 2022, https://counteringdisinformation.org/topics/platforms/complete-document-platforms.

33 Aliaksandr Herasimenka et al., 'Misinformation and Professional News on Largely Unmoderated Platforms: The Case of Telegram', Journal of Information Technology & Politics, 25 May 2022, 1–15, https://doi.org/10.1080/19331681.2022.2076272; Samantha Bradshaw, Renee DiResta, and Christopher Giles, 'How Unmoderated Platforms Became the Frontline for Russian Propaganda', Lawfare, 17 August 2022, https://www.lawfareblog.com/how-unmoderated-platforms-became-frontline-russian-propaganda-0.

34 EDMO and Institute for Data Democracy and Politics, 'Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access' (EDMO and Institute for Data Democracy and Politics, 2022), https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf.

35 Matthew Volk and Elizabeth Allendorf, 'Combating Political Disinformation with Online Political Ad Regulation' (Aspen Tech Policy Hub, 2020); Carah and Brodmerkel, 'Regulating Platforms' Algorithmic Brand Culture'

36 Nafia Chowdhury, 'Automated Content Moderation. A Primer', 2022

37 Francis Fukuyama, State Building: Governance and World Order in the 21st Century, (London: Profile Books Ltd., 2005).

38 Emma Llansó, 'CDT's Comments to Meta Oversight Board on Meta's Cross-Check Policy', Center for Democracy and Technology (blog), accessed 19 December 2022, https://cdt.org/insights/cdts-comments-to-meta-oversight-board-on-metas-cross-check-policy/; Gabriel Nicholas and Aliya Bhatia, 'Lost in Translation: Automated Content Analysis in Non-English Languages', Center for Democracy and Technology (blog), 18 August 2022, https://cdt.org/insights/lost-in-translation-automated-content-analysis-in-non-english-languages/.

39 Jeff Horwitz, 'Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. - WSJ', 2021, https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

40 Oversight Board, 'Policy Advisory Opinion on Meta's Cross-Check Program'.

over other users"[41]. In December 2022, Instagram said it would henceforth start to inform its users if their account was put under restrictions such as being banned from recommendation potential, a practice that had not been in place previously.[42]

A further issue is that platforms are often vague as to the basis of allowing exemptions from policy in terms of such broad criteria as "public interest" or "newsworthiness"[43]. YouTube says it operates a regime of partial exemptions from its policies for content on a case-by-case basis in terms of educational, documentary, scientific and aesthetic rationales.[44]

The breadth of these categories means that viewers and content creators are in the dark about why some 'content' makes the "cut" while other material attracts sanctions. Meta says "in rare cases" it will allow content that may violate policies "if it's newsworthy and if keeping it visible is in the public interest" and this outweighs risk of harm.[45] Yet, there is no clear elaboration of what constitutes public interest and only scant insight into its balancing test. Meta's "oversight board" has questioned the "newsworthiness" policy and its application in relation to some 90,7 million items removed under the Violent Graphic Content policy, saying that it was "unlikely that, over one year, only 17 pieces of content related to this policy should have been allowed to remain on the platform as newsworthy and in the public interest".[46] All this again shows that how policies stack up against each other is an area where more work is needed by the platforms.

On a related issue, YouTube offers an example of "borderline" content as "videos promoting a phoney miracle cure for a serious illness". But the platform also operates a policy whereby: "Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube.[47] This latter includes certain types of misinformation that can cause real-world harm, like promoting harmful remedies or treatments, ..."[48]. Such a policy mishmash is a recipe for poor and confusing enforcement of policy prescriptions.

Meta says "in rare cases" it will allow content that may violate policies "if it's newsworthy and if keeping it visible is in the public interest" and this outweighs risk of harm. Yet, there is no clear elaboration of what constitutes public interest and only scant insight into its balancing test.

41 Oversight Board, 'Oversight Board Publishes Transparency Report for Second Quarter of 2022 and Gains Ability to Apply Warning Screens | Oversight Board'.

42 Tom Gerken, Instagram has launched a new tool to let you know if your posts are barred from being recommended to other users, 2022, https://www.bbc.com/news/technology-63907699

43 Monika Bickert, 'Community Standards Enforcement Report, Second Quarter 2022', Meta (blog), 25 August 2022, https://about.fb.com/news/2022/08/community-standards-enforcement-report-q2-2022/; Ellen P. Goodman, 'Twitter's Newsworthiness Standard - What Is It?', Tech Policy Press, 7 December 2021, https://techpolicy.press/twitters-newsworthiness-standard-what-is-it/.

44 Youtube, 'How YouTube Evaluates Educational, Documentary, Scientific, and Artistic (EDSA) Content - YouTube Help', You tube, n.d, https://support.google.com/youtube/answer/6345162.

45 Meta, 'Our approach to newsworthy content UPDATED', 2022, https://transparency.fb.com/en-gb/features/approach-to-newsworthy-content/

46 Oversight Board, 'Oversight Board upholds Meta's decision in 'Sudan graphic video' case (2022-002-FB-MR)', 2022, https://www.oversightboard.com/news/1884013608451154-oversight-board-upholds-meta-s-decision-in-sudan-graphic-video-case-2022-002-fb-mr/

47 Youtube, 'Misinformation Policies - YouTube Help', Youtube Help, n.d., https://support.google.com/youtube/answer/10834785?hl=en.

48 Youtube Team, 'The Four Rs of Responsibility, Part 1: Removing Harmful Content', Youtube Official Blog, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/.

3 Implementation inadequacies

As outlined above, there are many issues with platforms' formal policies. But there is also much evidence of chasms between these policies and their application[49]. Much implementation is charged to "trust and safety" teams, which are often understaffed and subject to disbanding[50], and who frequently also have to manage and support inadequately trained and out-sourced content moderators who are cheaper for the platform than direct company employees[51].

This human contingent is complemented by external flaggers of problems and by fact-checkers. Algorithmic systems, informed by both policies and business imperatives, are supposed to do the heavy lifting on identifying and acting on the bulk of content that violates terms of service. This complexity of actors and factors compounds the challenges of enforcing policies through moderation of both specific items and whole categories of content.

> "Our policies are only as good as the strength and accuracy of our enforcement – and our enforcement isn't perfect.... One challenge is identifying potential violations of standards... Another challenge is accurately applying our policies to the content that has been flagged to us."
> Monika Bickert, Vice President, Meta (2018)[52]

A major factor underlying the presence of harmful content on the platforms is poor enforcement of policies. For example, research into anti-vaccine content on Facebook and Twitter between 1 February and 16 March 2021 showed that 65 percent of this content originated with 12 super-spreaders (the "Disinformation Dozen") with almost 60 million followers.[53] Another study found Facebook only containing, not decreasing, misinformation shared by "repeat offenders".[54] Actions by Facebook against anti-vaccination pages have been found to reduce engagement, but only temporarily, showing only partial success and the need for the company to operate ongoing vigilance and maintain capacity beyond peak periods of risk.[55] Analysis of one million posts on Telegram showed that this platform did not enforce terms of service that prohibit the promotion of violence.[56]

Many other reports exist of content remaining online that is problematic from a human rights vantage point, despite pledges by platforms to remove it.[57] In 2021, the Associated Press reported that hate speech

49 Ben Bradford et al., 'Report of the Facebook Data Transparency Advisory Group' (Data Transparency Advisory Group, 2019), https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.
50 Justin Hendrix, 'Facebook Whistleblower Frances Haugen and WSJ Reporter Jeff Horwitz Reflect One Year On', Tech Policy Press, 2 December 2022, https://techpolicy.press/facebook-whistleblower-frances-haugen-and-wsj-reporter-jeff-horwitz-reflect-one-year-on/.
51 Barrett, 'Who Moderates the Social Media Giants? A Call to End Outsourcing'.
52 Monika Bickert, 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', Meta (blog), 24 April 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards/.
53 Center for Countering Digital Hate, 'The Disinformation Dozen'.
54 Héloïse Théro and Emmanuel M. Vincent, 'Investigating Facebook's Interventions against Accounts That Repeatedly Share Misinformation', Information Processing & Management 59, no. 2 (March 2022): 102804, https://doi.org/10.1016/j.ipm.2021.102804.
55 David Broniatowski et al., Evaluating the Efficacy of Facebook's Vaccine Misinformation Content Removal Policies, 2022; Aoife Gallagher, Mackenzie Hart, and Ciarán O'Connor, 'Ill Advice: A Case Study in Facebook's Failure to Tackle COVID-19 Disinformation', 2021.
56 Jakob Guhl and Jacob Davey, 'A Safe Space to Hate: White Supremacist Mobilisation on Telegram', 2020.
57 CCDH, 'Star Framework. A Global Standard for Regulating Social Media', 2022; CCDH, 'Digital Hate. Social Media's Role In Amplifying Dangerous Lies About LGBTQ+ People', 2021; Ciarán O'Connor, 'Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok', 2021.

in Myanmar continued to thrive on Facebook, despite exposures of this three years earlier.[58] Hateful and extremist content is inconsistently removed on TikTok, according to one study.[59] Further research has shown that white nationalist content reaped millions of views on TikTok - despite the platform's rules against hateful expression.[60]

There are explanations for such continuing lapses:

A key factor is the problem of companies' monitoring of their own policy implementation of policy, in which they both set the metrics themselves and do their own evaluation.[61] (For a critical assessment of some metrics in play, see Part 1 in this series.) The result is that monitoring is generally self-serving and patchy. While many platforms record having set up mechanisms to implement commitments to freedom of expression[62], the adequacy of these mechanisms, and actual operation thereof, is impossible to assess externally without access to internal company data.[63] This reduces the credibility of the companies' own claims and weakens their accountability even to company shareholders. The known instances where platforms engage independent evaluators are so few as to stand out as exceptions.[64] Voluntary transparency reports produced by the bigger companies give some insight, but continue to lack data granularity. Platforms often give out statistics on numbers of items removed, and whether by automation, flagging, or government requests[65], but without the provision of more detail and context, it is hard to gauge the meaningfulness of such moderation. All this reflects a situation where most jurisdictions have no regulatory requirement for platforms to report on whether and how they are enforcing their policy commitments, or for them to avail sufficient data for independent assessment.

Limited features of voluntary self regulation allow problems to persist. The GNI operates a system of appointing independent assessors to evaluate member company performance in implementing the organisation's principles, with peer platforms making recommendations to improve compliance in any given case.[66] While this may help (with 11 companies due for assessment in 2021/2022)[67], and while a GNI member can be also be required to produce a "corrective action plan", there are no sanctions for poor performance. Further, the GNI process and documentation is confidential in character, with findings only made public in general form (notwithstanding Google's website stating "The GNI makes these company assessments publicly available"[68]). One such GNI report acknowledged further limitations noting that findings of compliance do not cover "... every decision to enter a market, or to develop, alter or acquire a product or service"[69].

Pre- and post- impact assessment is an area of noticeable weakness, with little being done by most companies, and poor transparency amongst the few who do.

All this comes back to the governance status quo where companies are free to choose if, how and when they might assess their performance, and the range of what should be assessed.

Most jurisdictions have no regulatory requirement for platforms to report on whether and how they are enforcing their policy commitments, or for them to avail sufficient data for independent assessment.

58 AP, '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.

59 Ciaran O'Connor, 'Gaming and Extremism. The Extreme Right on Twitch', 2021.

60 Alex Kaplan and Olivia Little, "'Groypers" Are Using TikTok to Promote White Nationalist Content, Evading the Platform's Ban', Media Matters for America, 2022, https://www.mediamatters.org/tiktok/groypers-are-using-tiktok-promote-white-nationalist-content-evading-platforms-ban.

61 Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', SSRN Scholarly Paper (Rochester, NY, 16 December 2019), https://papers.ssrn.com/abstract=3504930; Twitter, 'Removal Requests - Twitter Transparency Center', Transparency Twitter, 2021, https://transparency.twitter.com/en/reports/removal-requests.html#2021-jul-dec; Twitter, 'Rules Enforcement - Twitter Transparency Center', Transparency Twitter, 2021, https://transparency.twitter.com/en/reports/rules-enforcement.html; Bickert, 'Community Standards Enforcement Report, Second Quarter 2022'.

62 Ranking Digital Rights, 'The 2022 Ranking Digital Rights Big Tech Scorecard', 2022.

63 Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability', Internet Policy Review 9, no. 4 (18 December 2020), https://doi.org/10.14763/2020.4.1535.

64 Ernst & Young, 'Community Standards Enforcement Report Assessment Results', Meta (blog), 17 May 2022, https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/; Article One, 'Assessing the Human Rights Impact of Meta's Platforms in the Philipines' (Article One, 2020), https://about.fb.com/wp-content/uploads/2021/12/Meta-Philippines_HRIA_Executive-Summary_Dec-2021.pdf; Bradford et al., 'Report of the Facebook Data Transparency Advisory Group'.

65 Google, 'YouTube Community Guidelines Enforcement – Google Transparency Report', Transparency Report, 2022, https://transparencyreport.google.com/youtube-policy/removals.

66 GNI, 'Request for Proposals'.

67 GNI, 'Collaboration in Changing Times: GNI Annual Report 2021', 2021.

68 Google, 'Human Rights'.

69 GNI, 'Public Report on the Independent Assessment Process for Google, Microsoft, and Yahoo' (GNI, 2014), https://globalnetworkinitiative.org/wp-content/uploads/2016/10/GNI-Assessments-Public-Report.pdf.

Examples of other gaps between pronouncements and practice are regularly exposed by media and civil society:

There is systemically unequal treatment between categories of people[70], even although most policies are prima facie presented as being non-discriminatory (except in some cases, for example, where there are special dispensations for public figures).[71] Double standards have been exposed in moderating content which targets and emanates from non-dominant communities, including in cases of covert governmental influence operations.[72]

Platforms are accused in their policy implementation of being reluctant to upset ruling political forces, of not caring about smaller countries, or of deferring to bogus legal accusations.[73]

The roll-out of policies – for example, on elections – often falls short of pledges, and initiatives like creating elections operations centres are very uneven geographically.[74] For instance, Google has published transparency reports on political advertising, [75] but evidence for this practice in Africa could not be found.

There are intrinsic challenges in providing industrial-scale redress at the standard of due legal process[76]. Nevertheless, even as regards a minimum level of responsiveness by platforms to specific complaints by users and NGO watchdogs, platforms are widely criticised for being inadequate[77], even although some positive cases are reported.[78]

70 Marwa Fatafta, 'Meta's Clampdown on Palestine Speech Is Far from "Unintentional"', +972 Magazine, 9 October 2022, https://www.972mag.com/meta-arabic-palestine-censorship/; Sam Biddle April 13 2022 and 7:06 P.m, 'Facebook's Ukraine-Russia Moderation Rules Prompt Cries of Double Standard', The Intercept, 2022, https://theintercept.com/2022/04/13/facebook-ukraine-russia-moderation-double-standard/; Oversight Board, 'Policy Advisory Opinion on Meta's Cross-Check Program'.

71 Meta, 'Learn about WhatsApp's Fact-Checking Products, Partners and Investments | Meta Journalism Project', Fact-Checking on WhatsApp: Understanding Our Features, Products, Partners and Investments | Meta Journalism Project, 2022, https://www.facebook.com/journalismproject/fact-checking-whatsapp-products-partners.

72 Ángel Díaz and Laura Hecht-Felella, 'Double Standards in Social Media Content Moderation | Brennan Center for Justice' (Brennan Center for Justice, 2021), https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation; Julia Carrie Wong and Hannah Ellis-Petersen, 'Facebook Planned to Remove Fake Accounts in India – until It Realized a BJP Politician Was Involved', The Guardian, 15 April 2021, sec. Technology, https://www.theguardian.com/technology/2021/apr/15/facebook-india-bjp-fake-accounts; Julia Carie Wong, 'How Facebook Let Fake Engagement Distort Global Politics: A Whistleblower's Account | The Guardian – MediaWell', 2021, https://mediawell.ssrc.org/2021/04/15/how-facebook-let-fake-engagement-distort-global-politics-a-whistleblowers-account-the-guardian/.

73 Aman Abhishek, 'Overlooking the Political Economy in the Research on Propaganda', Harvard Kennedy School Misinformation Review, 1 April 2021, https://doi.org/10.37016/mr-2020-61; Daphne Keller, 'Balknization: Build Your Own Intermediary Liability Law: A Kit for Policy Wonks of All Ages', Balkinization (blog), 11 June 2019, https://balkin.blogspot.com/2019/06/build-your-own-intermediary-liability.html.

74 Accountable Tech, 'Facebook Is Failing Its Own Election Tests', 2020, https://accountabletech.org/research/facebook-is-failing-its-own-election-tests/.

75 Google, 'Political Advertising', Ads Transparency Google, n.d., https://adstransparency.google.com/political?political&region=KE.

76 Daphne Keller, 'The DSA's Industrial Model for Content Moderation – Verfassungsblog', 2022, https://verfassungsblog.de/dsa-industrial-model/.

77 Slate What Next TBD, Does Meta Even Care When Its Users Get Hacked?, 2023 https://slate.com/transcripts/cDE50jF5VW5mUXFscXBYOU5Bbk53L2dzSzFQQ091WjU0WWtoODM3RmJvbz0=

78 Bárbara Gomes Ribeiro et al., 'Analyzing the "Sleeping Giants" Activism Model in Brazil' (arXiv, 25 February 2022), http://arxiv.org/abs/2105.07523.

AUTOMATED MODERATION

The imbalances at stake concern both human and algorithmically-driven moderation in non-dominant languages.[79] Partly this reflects a skewed distribution in AI modelling, meaning that more than a billion people have minimal support.[80] These points have been brought out by whistleblowers at both Facebook and Twitter.[81]

Zuckerberg of Meta said in 2021 that the company uses "artificial intelligence to proactively report potentially problematic content to our team of reviewers, and in some cases to take action on the content automatically as well". He added further that "visual problems, like identifying nudity, are often easier than nuanced linguistic challenges, like hate speech. Our systems already proactively identify 96% of the nudity we take down, up from just close to zero a few years ago. We are also making progress on hate speech, now with 52% identified proactively."

But opacity surrounds such automated actions.[82] What is widely known, however, even by Meta engineers,[83] is the inability of contemporary Artificial Intelligence (AI) to recognise context, which is key to identifying potentially harmful content and when it may be approaching levels such as inciting and facilitating the organisation of efforts to violently overturn a legitimate election result. Limits of AI in moderation are also illustrated during one phase of the COVID-19 pandemic. At the time, YouTube increased reliance on AI for moderation, with the result that the number of removed videos escalated, as indeed did the number of human-decided reinstatements of this content on appeal.[84]

The bottom line is that there is an over-reliance by platforms on automated enforcement of their policies, even though the technologies are not adequate to this purpose, and this leads to over- and under-moderation in the enforcement of policies.[85]

Other identified differentials between policy and practice include:

Failure to moderate smartly, such as by filtering for alternate spellings of names of white supremacists on TikTok, leading to one particular video escaping the net and winning five million views and over 744 000 likes.[86] In Brazil, coded language inviting insurgents to call for the overturn of the country's 2022 election, such as "festa de selma" in multiple Telegram channels and on Twitter.[87] Meta and Google were reported to only begin removing calls for riots after a judicial order, while adverts calling for a military coup and allowed on Facebook scored 615 000 impressions.[88]

79 Nicholas and Bhatia, 'Lost in Translation'; Ryan Ryan-Mosley, 'The Internet Is Excluding Asian-Americans Who Don't Speak English', MIT Technology Review, 2021, https://www.technologyreview.com/2021/05/04/1024507/asian-american-language-justice-online-hmong/.

80 Courtney Radsch, 'AI and Disinformation: State-Aligned Information Operations and the Distortion of the Public Sphere', SSRN Electronic Journal, 2022, https://doi.org/10.2139/ssrn.4192038.

81 Justin Hendrix, 'Can Big Tech Platforms Operate Responsibly on a Global Scale?'; Hao, 'How Facebook and Google Fund Global Misinformation'; Karen Hao, 'The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why.', MIT Technology Review, 2021, https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

82 Justin Grandinetti, 'Examining Embedded Apparatuses of AI in Facebook and TikTok', AI & SOCIETY, 12 September 2021, https://doi.org/10.1007/s00146-021-01270-5.

83 Deepa Seetharaman, Jeff Horwitz and Justin Scheck, 'Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts', 2021, https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184?mod=e2tw

84 Rachel Kraus, 'YouTube Puts Human Content Moderators Back to Work | Mashable', Mashable, 2020, https://mashable.com/article/youtube-human-content-moderation; Issie Lapowsky, 'After Sending Content Moderators Home, YouTube Doubled Its Video Removals', Protocol, 25 August 2020, https://www.protocol.com/youtube-content-moderation-covid-19.

85 Oversight Board, 'Case Decision 2021-001-FB-FBR'. Paul Rottger et al.,' Hatecheck: Functional tests for hate speech detection models', (arXiv, 27 May 2021), https://arxiv.org/abs/2012.15606

86 Rieder and Hofmann, 'Towards Platform Observability'; Abbie Richards, 'TikTok Continues to Allow Videos of Neo-Nazi to Go Viral', Media Matters for America, 2022, https://www.mediamatters.org/tiktok/tiktok-continues-allow-videos-neo-nazi-go-viral.

87 Mark Scott, 'Digital Bridge: Lessons from Brazil — Surveillance capitalism — Digital Davos', 2023, https://www.politico.eu/newsletter/digital-bridge/lessons-from-brazil-surveillance-capitalism-digital-davos/

88 SumOfUs, 'URGENT REPORT. How Meta and Google enabled and profited from the terrorist attacks in Brazil's capital', 2023, https://s3.amazonaws.com/s3.sumofus.org/images/Research_SumOfUs_Brazilian_Riots_January_11th_2023.pdf; Hannah Gelbart, Juliana Gragnani and Ricardo Senra, Brazil: The code word used to invite protesters to a riot, 2023, https://www.bbc.com/news/blogs-trending-64223574

While investment in fact-checking has been shown empirically to debunk various false claims[89], the way such verification operates and how and when it is deployed (or not) lacks transparency. Some evidence also shows that its impact is limited.[90] The effectiveness of warnings (and of labels[91]) has also been assessed as unclear.[92] This is also evident in a review of 42 scientific papers covering the role of accuracy prompts, debunking, friction, inoculation, lateral reading, media-literacy tips, rebuttals of science denialism, self-reflection tools, social norms, and warning and fact-checking labels.[93] The lack of platform disclosure hinders assessment of how these and other moderation-linked measures may be meaningful in terms of achieving policy objectives.

The lack of platform disclosure hinders assessment of how these and other moderation-linked measures may be meaningful in terms of achieving policy objectives.

Content removal and deplatforming of individuals and organisations are dramatic sanctions which call out for full observance of due process and appeal systems by the platforms using such measures. Where such conditions are met, these steps have immediate and evident effect on the platform concerned, and they show a rights-respecting role by the company concerned. But there are still problems of "whack-a-mole" where the targets pop up again, and also of migration of affected actors to platforms with weaker standards. However, even with such strong actioning by a given platform, much problematic content and many actors seeking to violate human rights on that same service still escape detection let alone moderation. There are also false positives whereby legitimate expression and actors (eg. linked to journalistic reportage) are wrongly penalised.[94] It is not possible, in the absence of data, to verify platform claims and assess whether they could be doing more.

89 Nathan Walter et al., 'Fact-Checking: A Meta-Analysis of What Works and for Whom', Political Communication 37, no. 3 (3 May 2020): 350–75, https://doi.org/10.1080/10584609.2019.1668894.

90 Man-pui Sally Chan et al., 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation', Psychological Science 28, no. 11 (November 2017): 1531–46, https://doi.org/10.1177/0956797617714579; Frederico Batista Pereira et al., 'Fake News, Fact Checking, and Partisanship: The Resilience of Rumors in the 2018 Brazilian Elections', The Journal of Politics 84, no. 4 (October 2022): 2188–2201, https://doi.org/10.1086/719419.

91 Kevin Aslett et al., 'News Credibility Labels Have Limited Average Effects on News Diet Quality and Fail to Reduce Misperceptions', Science Advances 8, no. 18 (6 May 2022): eabl3844, https://doi.org/10.1126/sciadv.abl3844.

92 Björn Ross and Anna-Katharina Jung, 'Fake News on Social Media: The (In)Effectiveness of Warning Messages', International Conference on Information Systems (ICIS), 1 January 2018, https://www.academia.edu/38296120/Fake_News_on_Social_Media_The_In_Effectiveness_of_Warning_Messages.

93 Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan Herzog, Ullrich Ecker, Stephan Lewandowsky and Ralph Hertwig, Toolbox of interventions against online misinformation and manipulation, n.d. https://interventionstoolbox.mpib-berlin.mpg.de/index.html

94 Evelyn Douek, 'The Rise of Content Cartels' (Knight First Amendment Institute at Columbia University, 2020), http://knightcolumbia.org/content/the-rise-of-content-cartels.

4 Pointers towards possible improvement

In regard to mechanisms for policy implementation, there is generally an absence of institutional frameworks for inter-industry co-operation. An exception is the Global Internet Forum to Counter Terrorism. However, it lacks the participation of smaller (but significant) platforms even though fostering violent extremism can ultimately affect everyone.[95] There are limited experiences of institutionalised external stakeholder involvement in regard to problematic content issues beyond some giant platforms recognising certain individuals or entities as "trusted flaggers" which are criticised for lacking explicit rationale or elaborated role.[96]

Meta has set up an "oversight board" which, despite its title, can decide against the company on only specific cases and merely advise it on more general policy matters. While these are steps in a direction that can lead to better performance by the platforms, there are some backward moves as well. Notably, Twitter disbanded its trust and safety advisory group in December 2022.[97] Its CEO Elon Musk also decided that signals for muting and blocking content generally would henceforth take account of the patterns of paying subscribers on the platform, raising fears that this could serve those actors wanting to weaponise adversarial reporting possibilities.[98]

The lack of institutionalised engagement by platforms with users and other multiple stakeholders policy operations is further apparent in the phenomenon of geographically uneven or non-existent relationships with election management bodies, media and civil society, although some experiences are cited in Part 3 of this series.[99]

On the more positive side, there is a recent case of Meta offering to share with smaller platforms its technology to moderate against visual content deemed to be violating.[100] Google has also announced a similar initiative.[101]

Further, there are increasing reports by platforms of their individual enforcement of policy proscriptions against influence operations,[102] and also reports of inter-company co-operation in this regard.[103] Facebook said in 2022 that it had taken down more than 200 covert influence operations since 2017 in nearly 70 countries that operated in more than 40 languages[104].

---

95 Man-pui Sally Chan et al., 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation', Psychological Science 28, no. 11 (November 2017): 1531–46.

96 Naomi Appelman and Paddy Leerssen, 'On "Trusted" Flaggers', n.d.

97 Matt O'Brien and Barbara Ortutay, 'Musk's Twitter Disbands Its Trust and Safety Advisory Group | AP News', 2022, https://apnews.com/article/elon-musk-twitter-inc-technology-business-a9b795e8050de12319b82b5dd7118cd7.

98 Alex Stamos, 'The speedrun through "every trust and safety problem others have experienced over the last decade but that I don't know anything about" continues. Up next, adversarial reporting', 18 December 2022, https://cybervillains.com/@alex

99 Robert Krimmer et al., 'Elections in Digital Times: A Guide for Electoral Practitioners - UNESCO Digital Library', UNESCO Digital Library, 2022, https://unesdoc.unesco.org/ark:/48223/pf0000382102.

100 Nick Clegg, 'Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO', Meta (blog), 13 December 2022, https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/.

101 Tech Against Terrorism to Build Content Moderation Tool with Google Jigsaw, 2023, https://www.techagainstterrorism.org/2023/01/09/tech-against-terrorism-to-build-content-moderation-tool-with-google-jigsaw/

102 Nathaniel Gleicher, 'Coordinated Inauthentic Behavior Explained', Meta (blog), 6 December 2018, https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/; Alliance for Securing Democracy and Graphika, 'IOA - Information Operations Archive', IOA, n.d., https://www.io-archive.org/#/; Graphika and Stanford Internet Observatory, 'Unheard Voice. Evaluating Five Years of pro-Western Covert Influence Operations', Graphika, 2022, https://graphika.com/reports/unheard-voice; Grossman, Gallagher, and Johnson-Kanu, '#ZakzakyLifeMatters: An Investigation into a Facebook Operation Linked to the Islamic Movement in Nigeria'; Shelby Grossman, Sean Gallagher, and Ada Johnson-Kanu, 'One Face, Many Names: An Investigation into Fake NGOs and Media Outlets Linked to Harouna Douamba on and off Facebook', n.d.; Nimmo, 'Meta's Adversarial Threat Report, Second Quarter 2022 | Meta'.

103 Gleicher, 'Coordinated Inauthentic Behavior Explained'.

104 Guy Rosen and Nathaniel Gleicher, 'Protecting People From Online Threats In 2022', 2022. https://about.fb.com/news/2022/12/protecting-people-from-online-threats-in-2022

Nevertheless, undetected cases of manipulation driving potentially harmful content are still being found by external observers, suggesting that there are limits to algorithmic detection of what Meta calls "co-ordinated inauthentic behaviour".[105]

Another positive development has been companies' policies and practices to proactively provide authoritative health or electoral information, although less is generally being done in relation to foregrounding of credible news sources, and the criteria for ongoing eligibility to be regarded as such are open to critique. The Council of Europe's Guidance Note on the Prioritisation of Public Interest Content Online cautions that "regimes of prominence" for such content curation should be transparent about the criteria for what counts as public interest content, the selection process, and its actual consumption by audiences.[106] The Note further encourages States "to develop – in cooperation with each other, with platforms and civil society organisations – a co-regulatory framework or other appropriate and proportionate forms of risk-based governance". This mechanism would ensure adequate and independent democratic oversight of algorithmic systems, especially with respect to access, distribution, and prioritisation of content. "Such oversight should include reporting duties on algorithmic content curation and prioritisation to the independent media and/or platform regulatory authorities or other designated bodies entrusted with maintaining and promoting pluralism and diversity in the public sphere".[107]

Notwithstanding extensive efforts at implementation, the bottom line appears to be that the problems of misinformation, disinformation and hate speech continue on many platforms, and the problem is not reducible to simple quantitative prevalence metrics. The point is that such expression potentially harms human rights, while also diverting from, and often seeking to discredit, information as a public good, and that it persists despite extant policies and practices. Without external pressure from public opinion and diverse regulatory arrangements, platforms seem unlikely on their own accord to introduce and pay for more profound changes.[108]

Without external pressure from public opinion and diverse regulatory arrangements, platforms seem unlikely on their own accord to introduce and pay for more profound changes.

105 Zoé Fourel and Cooper Gatewood, 'Amplifying Far-Right Voices: A Case Study on Inauthentic Tactics Used by the Eric Zemmour Campaign', ISD, accessed 21 December 2022, https://www.isdglobal.org/isd-publications/amplifying-far-right-voices-a-case-study-on-inauthentic-tactics-used-by-the-eric-zemmour-campaign/.

106 CDMSI, 'Guidance Note on the Prioritisation of Public Interest Content Online' (Council of Europe, 2021), https://rm.coe.int/cdmsi-2021-009-guidance-note-on-the-prioritisation-of-pi-content-e-ado/1680a524c4.

107 CDMSI.

108 Christopher T. Marsden, Ian Brown, and Michael Veale, 'Responding to Disinformation: Ten Recommendations for Regulatory Action and Forbearance', preprint (LawArXiv, 13 November 2020), https://doi.org/10.31228/osf.io/aerw9.

5 Regional imbalance and neglect of local languages in content management

Analyst Evelyn Douek has argued: "You cannot enter a market without the language understanding or the contextual understanding or political expertise and expect this kind of moderation to be sufficient or prevent harm." [109]

Evidence of the risk of harms is illustrated by UNESCO's project Social Media 4 Peace, which has surfaced how platforms have grave deficits in investment in treating potentially problematic content in the countries in which the project is being implemented (Bosnia-Herzegovina, Colombia, Indonesia and Kenya).[110] Similarly, Meta's own "oversight board" has found that the company has a mismatch between the volume of content that is designated for enhanced review and the inadequate human resources allocated to the task.[111]

There is extensive geographical variation in terms of how platform influence on content and associated communications is aligned to national laws and policies, with such contours being more closely followed in countries and regions that constitute major business markets. Evidence shows the big platforms are generally lacking in regard to linguistically-appropriate moderation services and redress channels. This is even in regard to the provision of their primary guidelines document in significant languages.[112]

The imbalance in language investments means that harms to human rights are potentially less prevented or mitigated in underserved countries and languages. Whistleblower Frances Haugen revealed that in 2020, Facebook devoted 13% of its budget for developing algorithms to detect misinformation to regions outside the US, which raises the issue of the value of advance impact assessments to evaluate what percentage would be appropriate to different contexts.[113] The EU's 2022 Digital Services Act may lead platforms to operate more consistently across Europe.

META AND LANGUAGE

The world's biggest platform has said, without giving detail, that its content review teams operate in over 20 sites and cover more than 70 languages.[114] Further disclosure is conspicuously lacking, meaning that it is not possible for outsiders to gauge the actual extent of investment in language competencies and with what impact on content moderation. However, confirming its interest in news in multiple tongues, Meta in 2022 was offering registration for news providers on an index covering 42 languages.[115]

In July 2022, Zuckerberg announced the open-sourcing of an AI model that can translate across 200 different languages, which he said would play a part in relation to content moderation[116]. However, since machine learning depends on data, and there is a grave linguistic imbalance in this regard, questions can be asked about effectiveness of Meta's AI optimism. A further difficulty here will be whether the technology can be developed to recognise new and coded language, imagery and emoticons.

109 Zecharias Zelalem, 'Why Facebook Keeps Failing in Ethiopia', Rest of World, 13 November 2021, https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia/.
110 Article 19, 'Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society', 2022, https://www.article19.org/wp-content/uploads/2022/06/Summary-report-social-media-for-peace.pdf.
111 Owens, 'Social-Media Platforms Failing to Tackle Abuse of Scientists'.
112 Washington Post, 'Why Facebook Won't Let You Control Your Own News Feed'; Oversight Board, 'Oversight Board Submission to the Special Rapporteur on Freedom of Opinion and Expression: Challenges in Times of Conflicts and Disturbances'.
113 Owens, 'Social-Media Platforms Failing to Tackle Abuse of Scientists'.
114 Meta, 'Corporate Human Rights Policy' (Meta, 2021), https://about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf.
115 See for example, https://www.facebook.com/business/help/316333835842972?id=644465919618833
116 Mark Zuckerberg, Facebook, 2022, https://www.facebook.com/zuck.

Meta's "oversight board" meanwhile has recorded that the company has committed to "translate its rules into languages spoken by 400+ million people". By UNESCO's count in December 2022 of the more than 70 language versions of the company's "community standards"[117], the target is more than reached. At the same time, this step on its own does not translate into better geographical inclusiveness. UNESCO's analysis of seven Transparency Reports from 2020 to mid-2022 by Meta's "oversight board" show that almost 70% of the close to two million complaints they received come from the US and Canada, and Europe, while regions with high user numbers but using other languages are not taking their concerns to this company channel.

Furthermore, beyond translations of policies and rules, there is the issue of enforcement. The "oversight board" itself acknowledges "concerns about whether Meta has invested sufficient resources in moderating content in languages other than English".[118] The board signals additionally that this goes down to the issue of language in automated notifications to users about why their content appears to have violated a rule. It further addresses the language issue in regard to human review of possibly violating content from the elite who are given special treatment under the company's cross-check policy. In this regard, it states that "the Early Response Team does not have language or regional expertise and it relies on translations and contextual information provided by the relevant Regional Market Team to assess the content". Additionally, the board has opined that instead of focusing on contexts with greater risks to human rights, including freedom of expression,[119] it appears that: "Meta does not prioritise training its automated processes on less-spoken languages and less lucrative markets. Limited investment in moderation in these languages limits the ability of algorithms to identify topics in such content. This suggests that users in these markets, including the Global South, may be disadvantaged..."[120]

For its part, Meta argues that it does prioritise for countries at risk, including in relation to elections, and that in a crisis it determines what support and teams should be dedicated to a particular country or language.[121] The potential for offline harm and violence in these cases would thus trigger the need to increase company capacity in relevant languages for such contexts.[122] Evidently, systematic prior risk-assessments would be essential in such cases, since trained and experienced content moderators in the relevant languages need to be recruited and prepared well in advance – keeping in mind also that this function is largely outsourced. At present, it is in the companies' own prerogative about whether to conduct such advanced assessments.

However, although the EU's Global Data Protection Regulation has wide international impact, it can be noted that entities such as WhatsApp still operate weaker privacy standards outside the EU's member countries. It is therefore not a given that geographical imbalances in content moderation around the world will diminish as a result of EU standard-setting for platforms within its jurisdiction.

Numerous calls have been made for platforms to go beyond language and also employ more local policy, political and cultural expertise.[123] Spending additionally is needed in languages that use different scripts to those that are currently served[124].

117 See for example https://transparency.fb.com/en-gb/policies/community-standards/
118 Meta, 'Learn about WhatsApp's Fact-Checking Products, Partners and Investments | Meta Journalism Project'.
119 Oversight Board, 'Policy Advisory Opinion on Meta's Cross-Check Program'.
120 Oversight Board. See also: Cat Zakrzewski, Gerrit De Vynck, Niha Masih and Shibani Mahtani, 'How Facebook neglected the rest of the world, fueling hate speech and violence in India', 2021, https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/
121 Miranda Sissons, 'An Independent Assessment of Meta's Human Rights Impact in the Philippines', Meta (blog), 2 December 2021, https://about.fb.com/news/2021/12/metas-human-rights-work-philippines/.
122 Miranda Sissons and Nicole Isaac, 'Our Approach to Maintaining a Safe Online Environment in Countries at Risk', Meta (blog), 23 October 2021, https://about.fb.com/news/2021/10/approach-to-countries-at-risk/.
123 For example, Aim Sinpeng et al., 'Facebook: Regulating Hate Speech in the Asia Pacific' (University of Sidney and University of Queensland, 2021).
124 See Yudhanjaya Wijeratne, Nisansa de Silva, Yshothara Shanmugarasa (2019) Natural Language Processing for Government: Problems and Potential, ResearchGate

6 Recommendations

- Statutory authorities should not seek to take over the direct policy formulation nor the ongoing moderation work by the companies themselves, but they can set objectives, policy standards and process benchmarks that apply to solo-, self- and co-regulatory mechanisms which can ensure more effective performance by the platforms themselves.

- It can be strongly encouraged that an array of regulatory arrangements combined can work to get moderation policies and practices properly integrated with international human rights standards, and that these should also spell out how to balance between global and local dimensions.

- Guidance could insist that all relevant platform policy documents should be public, and in the primary languages where companies avail their services.

- It could also require platforms to elaborate on how they decide what will take priority when there are competing policy provisions or when policy and business models pull in different directions.

- Platform companies can be compelled to undertake appropriate scenario planning and implement due diligence exercises into the full range of risks anticipated in upcoming trends implicating content and communications, and to provide detail on how they will mitigate these through policy and implementation measures.

- Legal regulation for wide-ranging transparency can help improve the credibility of companies about their performance claims.

- There should be requirements for independent assessment of metrics in terms of how policy implementation is working and provide for independent monitoring and auditing of policy implementation.

- Guidance can be given about assessing the balances between AI and humans in content moderation, and require that there be improved channels for appeal when automated moderation takes place.

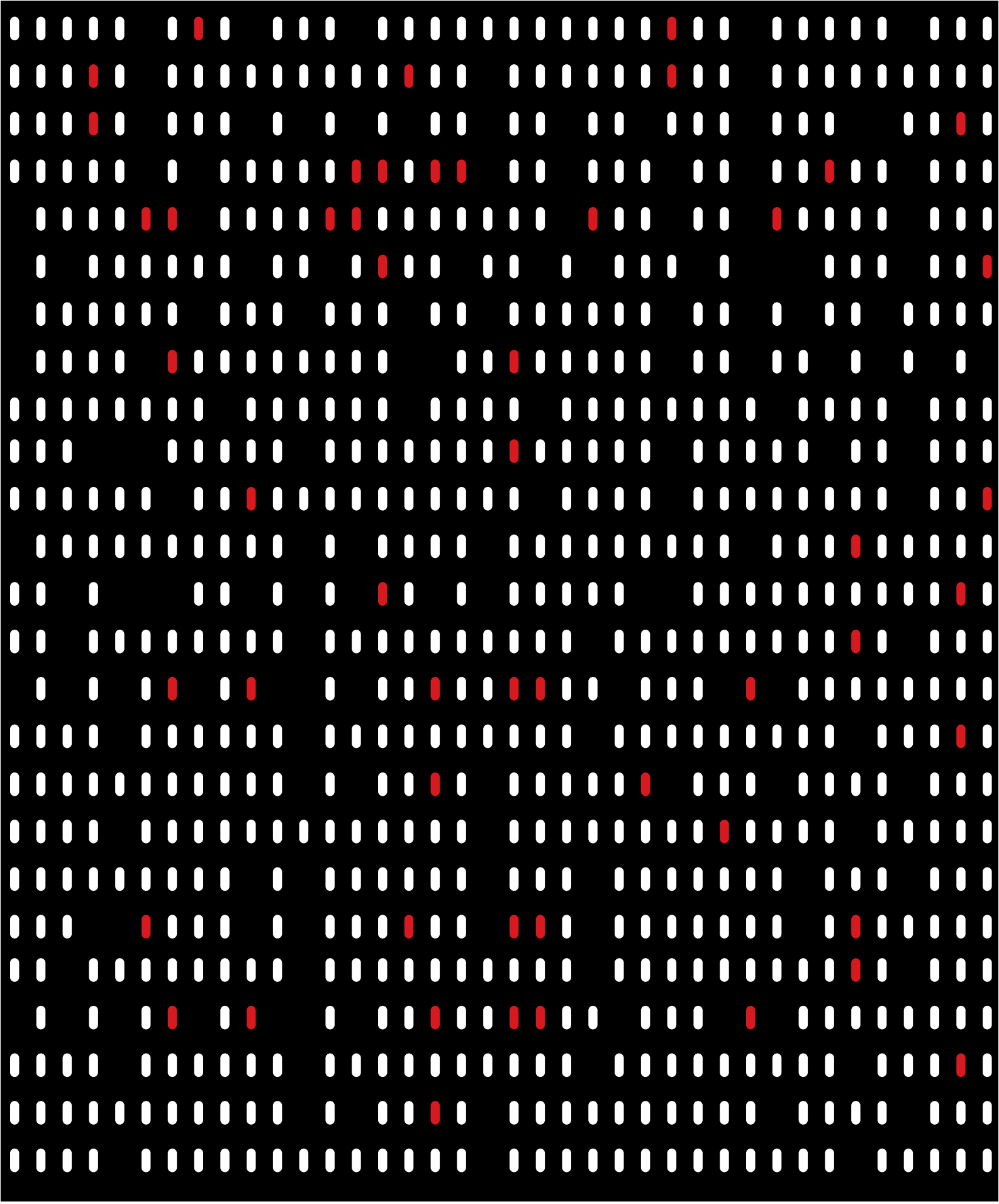
- Regulation can require companies to respect the right to equality by addressing inequalities in both policies and implementation.

7 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?
- In what ways can policies (including those against bullying, spam and scams) be mapped to different human rights?
- How can pluralistic company policies and practices be linked to wider shared standards operated by institutions of self-regulation like social media councils?
- How might multi-stakeholder participation and co-operation be structured into policy development, monitoring, assessment and revision?
- If a platform cannot provide reasonable moderation or remedial presence in a given jurisdiction, how might regulators assess whether or not it should still offer services there bearing in mind that such services may also have benefits for free expression and access to information?
- What distinctions might be drawn between content policies and enforcement by different sizes and kinds of platforms, including centralised and decentralised ones?

Comments can be sent to e-mail: internetconference@unesco.org with the subject line: Response to draft background paper or on the link here: https://forms.gle/qyeKP7gqjztujpgv6



This research was supported by UNESCO