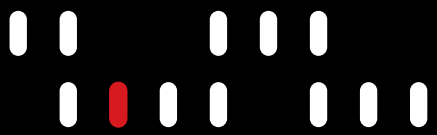
PART 3

DIGITAL PLATFORM GOVERNANCE AND THE CHALLENGES FOR TRUST AND SAFETY

RESEARCH ICT AFRICA

POSSIBLE REGULATORY SOLUTIONS TO ADDRESS CONCERNS WITH THE PLATFORMS



KEY TRENDS UNCOVERED

- Platform problems are linked to the fact that they are not self-governing according to agreed industry standards but mainly 'solo-governing' when it comes to content curation and moderation.
- Reaction to the failure of current platform efforts to regulate content includes the danger of over-regulation by state entities, which carries real risks to freedom of expression.
- The purview of what may need to be part of new regulatory arrangements includes the interplay between policy, practice, business models and technology.
- There is a pluralism of platforms and other actors in the "tech stack", who have different roles in the online content landscape, with concomitant implications for regulatory arrangements.
- Independent media, whistle-blowers and civil society organisations are significant factors in pushing platform accountability but mechanisms of transparency should be considered for regulatory protections and support.
- New technology is raising new challenges for platforms' content moderation.
- Platform policy and practice is especially significant for elections.

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 3 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.

1 Context

Platform companies– big and small – have varying reaches and ranges of services (eg. search, communication, content production and consumption, commerce, etc.).[1] Where they are implicated in content issues, this makes them central to UNESCO's interests in "Information as a public good" which is threatened by online disinformation, misinformation and hate speech.[2]

Globally, it is evident that a major portion of online communications occurs via a handful of digital platforms, listed on major stock exchanges. This reality has led researcher Alex Kasodomski to argue that "the infrastructure underlying the public commons is in the hands of a few powerful corporations and their shareholders", adding that "Governments must look to the future with regulation which fits distributed platforms, greater participation by democracies in standards setting, and investment in a new direction for technology".[3]

The scale of the issue is evident from statistics announced by Datareportal in October 2022, which put the figure of users of social media worldwide at 4,7 billion. This represents more than 75% of the world's adult population (although the aggregate conceals differences among sexes and among countries).[4] In January 2022, Datareportal identified Meta's acquisition – WhatsApp – as the world's favourite communications service, followed by the company's Instagram and Facebook, and then Wechat and Douyin. TikTok was reported as the most-downloaded mobile app in 2021, followed by Instagram, Facebook, Whatsapp and Telegram.[5] Facebook itself reported 3.71 billion monthly active people in September 2022.[6]

This massive penetration into societies has not just happened organically – it results from deliberate business strategies in this particular sector which put a premium on rapid growth (over other considerations like safety) in order to please investors.[7] The scale issue calls for particular attention to these entities, as is reflected for example in the Digital Services Act (DSA) of the European Union (EU), which recognises "very large" operators as a distinct category meriting the most regulation. Another perspective worth noting is the Council of Europe observation that "(t)he responsibility of intermediaries to respect human rights and to employ adequate measures applies regardless of their sise, sector, operational context, ownership structure or nature. The scale and complexity of the means through which intermediaries meet their responsibilities may vary, however, taking into account the severity of impact on human rights that their services may have."[8]

Concerns with major operators should not overshadow that the issues are also germane to much small services which can also have significant roles in regard to content that can harm human rights. There is not space to elaborate the diverse regulatory arrangements suited to different size platforms. Instead the focus that follows below on the need for minimum standards of all platform operators that can improve protection of human rights across the board and be designed to ensure that severe adverse impacts can be anticipated and mitigated in all cases.

---

1 Sarah Hartmann, 'Policy Developments in the USA to Address Platform Information Disorders' (Perspectives on Platform Regulation, Nomos Verlagsgesellschaft mbH & Co. KG, 2021), 99–118, https://doi.org/10.5771/9783748929789-99.

2 "UNESCO General Conference Endorses the Windhoek +30 Declaration", 24 November, 2021. https://www.unesco.org/en/articles/unesco-general-conference-endorses-windhoek-30-declaration

3 Ellen Judson et al., 'THE CONTOURS OF STATE-ALIGNED GENDERED DISINFORMATION ONLINE', 2020.

4 Simon Kemp, 'The Global State of Digital in October 2022 — DataReportal – Global Digital Insights', 2022, https://datareportal.com/reports/digital-2022-october-global-statshot.

5 Simon Kemp, 'Digital 2022: Global Overview Report', DataReportal – Global Digital Insights, 2022, https://datareportal.com/reports/digital-2022-global-overview-report.

6 'Meta's Adversarial Threat Report, Third Quarter 2022', *Meta* (blog), 22 November 2022, https://about.fb.com/news/2022/11/metas-adversarial-threat-report-q3-2022/.

7 Paul Barrett M, Justin Hendrix, J, and Grant Sims, 'Fueling the Fire: How Social Media Intensifies U.S. Political Polarization — And What Can Be Done About It', 2021, https://bhr.stern.nyu.edu/polarization-report-page.

8 Council of Europe, 'Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries', 2018, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14. See also: Jack Hardinges and Jared Robert Keller, 'What Are 'bottom-up Data Institutions and How Do They Empower People?', *What Are 'bottom-up Data Institutions and How Do They Empower People?* (blog), n.d., https://theodi.org/article/what-are-bottom-up-data-institutions-and-how-do-they-empower-people.

It is the case that as per the Australian eSafety Commissioner "one sise does not fit all" providers, who differ in terms of resources, technical architecture and user base.[9] It is also be the case that a single super-regulator cannot do justice to platform governance. For instance on some issues, a mix of self- and co-regulatory arrangements may be appropriate (such as on aspects of transparency reporting); on others a multistakeholder mechanism may be better (like on identifying and promoting authoritative content). In this context, the value of an overarching regulatory guidance is that it can interpret what it means to have an approach centred on human rights as foundational to all regulatory possibilities, with a minimum requirements relevant at least to the bulk of platforms. Further, as covered in Part 2 in this series, local context and related balances of rights[10] are significant issues in platform regulatory arrangements. In this light, overarching regulatory guidance based on international human rights benchmarks can work to facilitate cross-jurisdictional harmony.[11]

It is in taking such an approach that a clear cost-benefit analysis becomes possible of the array of possible regulatory initiatives. These should seek to minimise compliance costs while assessing benefits to society, businesses, democracy, sustainable development and human rights, and inform an optimum balancing in the ongoing calculus. This cost-benefit issue further points in the direction of considering alternatives to adopting new legal regulatory steps[12] – and applying such alternatives to certain purposes and sites. Examples are boosting media and information literacy, incentivising different platform business models, making use of existing legal regulations, and improving the enforcement thereof.

Of relevance to regulation is that the literature also shows that the particular attention-economics and micro-targeting advertising models, and spending priorities, of the major platforms are implicated as major determinants factors in accounting for content of at least potential negative impact for human rights. Researchers Luca Belli and Nicolo Zingales observe: "Due to their for-profit nature, it is also logical to expect that the design and implementation of platforms' dispute resolution mechanisms will tend to be largely driven by considerations of cost minimisation and avoidance of potential liability, rather than the maximisation of the scope of protection of individual rights".[13] However, that many companies in capitalist markets operate with similar logics does not preclude new regulatory arrangements that could counter the specific negative outcomes in this sector.

"Due to their for-profit nature, it is also logical to expect that the design and implementation of platforms' dispute resolution mechanisms will tend to be largely driven by considerations of cost minimisation and avoidance of potential liability, rather than the maximisation of the scope of protection of individual rights".

The challenges of platforms' potential content harm to human rights are also evident in regard to services provided by companies like Paypal, eBay and Etsy. Actors at other levels of the "tech stack" also becoming impactful content-related gatekeepers.[14] Examples are internet access providers, cloud security providers, and app stores which, through both omission and commission, can play decisive parts in what content flows online and what does not. For example, some decide whether a given service they mediate has what they consider adequate moderation in place.[15] Online funding services are also accused of "bankrolling bigotry" due to an absence of moderation.[16] All this has prompted calls for a holistic overall policy approach that ranges from social media through to search, cloud service providers and e-commerce services, and accommodates a range of tiered regulatory arrangements within this bigger picture.[17]

At the same time, alternative regulatory possibilities (eg. with more place for multistakeholder and user regulatory modalities) may be appropriate in regard to non-profit platforms such as Wikipedia and to the application of protocols enabling the "fediverse" (including for example, the Mastodon network). These entities have shown themselves to operate effective decentralised moderation systems, and be less vulnerable to potential content harms[18]. They exist alongside other models like Slashdot.org and Reddit which are more user driven rather than top-down in terms of content issues.

9 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

10 Colin J. Bennett and David Lyon, 'Data-Driven Elections: Implications and Challenges for Democratic Societies', *Internet Policy Review* 8, no. 4 (31 December 2019), https://doi.org/10.14763/2019.4.1433.

11 Internet and Jurisdiction Policy Network, 'Toolkit: Cross border moderation', 2021

12 UK National Audit Office, 'Using alternatives to regulation to achieve policy objectives. Summary.', 2014, https://www.nao.org.uk/wp-content/uploads/2014/06/Using-alternatives-to-regulation-to-achieve-policy-objectives-summary.pdf

13 Luca Belli and Nicolo Zingales, 'Platform Value(s): A Multidimensional Framework for Online Responsibility', *Computer Law & Security Review* 36 (1 April 2020): 105364, https://doi.org/10.1016/j.clsr.2019.105364.

14 GDI, 'The Quarter Billion Dollar Question: How Is Disinformation Gaming Ad Tech?', 2019, https://www.disinformationindex.org/.

15 Travis Clark, 'Trump's Truth Social App Is Officially on the Google Play Store after Complying with Moderation Rules', Business Insider, 2022, https://www.businessinsider.com/trump-truth-social-android-app-google-play-store-2022-10; Karissa Bell, 'Apple Threatens to Ban Parler from App Store', Engadget, 2021, https://www.engadget.com/apple-threatens-parler-ban-000548898.html.

16 GDI, 'The Business of Hate: Bankrolling Bigotry in Germany and the Online Funding of Hate Groups', 2021, https://www.disinformationindex.org/.

17 GDI, 'Disrupting Online Harms: A New Approach' (Global Disinformation Index, 2021).

18 Alan Z. Rozenshtein, 'Moderating the Fediverse: Content Moderation on Distributed Social Media', Journal of Free Speech Law, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213674

2 The landscape of platform governance

The prevalent governance regime for most platforms across issues of hate expression, misinformation and disinformation, is one of "solo-regulation"[19] or "private ordering initiatives"[20] by individual companies. There is nevertheless a degree of regulatory-compelled conformity across the platform industry, such as with the USA's copyright laws and EU's data protection standards.

In many countries, platforms also operate within the frame of national e-commerce and cybersecurity regulation, but such rules are generally inappropriate for online content issues (although they are also often too broadly worded and misapplied to cases of expression).

Unlike the common case with the news media industry and advertising industry, sector–wide self-regulatory[21] systems across diverse platform companies are generally non-existent. Some initiatives that involve cross industry players include the Global Network Initiative, the Digital Trust and Safety Partnership, and the Global Internet Forum to Counter Terrorism, and the Content Authenticity Initiative. However, these arrangements focus more on encouraging 'good practices' than on acting as joint governance mechanisms for how platforms deal with problematic content.

At the level of co-regulation, the governance landscape for platforms in regard to content includes relatively recent cases notably in the EU and Australia in relation to disinformation and hate speech. In these systems, state actors work with industry to set standards, which may be voluntarily or legally binding, but without setting up a full official regulatory apparatus for implementation at operational level. There are varying degrees of statutory authorisation of these arrangements, with some being less formal than others.

Concerning multi-stakeholder regulatory models, the company Meta operates a limited form by engaging a selection of civil society individuals in what it calls its "oversight board". This mechanism has authority in regard to specific cases of content being taken down or left up, but not on the more fundamental policy issues. Several criteria for assessing such multi-stakeholder models have been proposed as prerequisites to qualify as genuine mechanisms, highlighting issues around the participant selection process, independence of operations, and the remit in terms of authority.[22]

Other than the above landscape, the governance of the platforms in regard to disinformation, misinformation and hate speech largely remains one of laissez-faire for each entity, although there are cases of partial, patchy and fragmented regulation by a range of legal authorities. This setup fundamentally enables "governance by platforms"[23] which, combined with their business models and of growth and profit, has been a factor in the rise of systemic content-related risks to human rights.

In this context, there is an increasing trend towards regulation in the form of new legal measures by state bodies. Meta has noted that content-related legislation "has required us in the past, and may require us in the future, to change our products or business practices, increase our costs, or otherwise impact our operations or our ability to provide services in certain geographies."[24]

---

19 The concept is based on Marko Milosavljević and Sarah Broughton Micova, 'Banning, Blocking and Boosting: Twitter's Solo-Regulation of Expression', *Medijske Studije* 7, no. 13 (2016): 43–58, https://doi.org/10.20901/ms.7.13.3.

20 Council of Europe, 'Recommendation CM/Rec (2022)11 of the Committee of Ministers to Member States on Principles for Media and Communication Governance', 2022, https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680a61712.

21 Article 19, 'The Social Media Councils: Consultation Paper', 2019, https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf

22 Jyoti Panday, Milton Mueller and Farzaneh Badii, 'Multistakeholderism and Platform Content Governance: An Assessment Framework with Applications', 2022 https://www.internetgovernance.org/wp-content/uploads/MS-Content.docx-1.pdf

23 Robert Gorwa, 'What Is Platform Governance?', *Information, Communication & Society* 22, no. 6 (12 May 2019): 854–71, https://doi.org/10.1080/1369118X.2019.1573914.

24 Meta Platforms, Inc., 'Annual Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act 1934 for the Fiscal Year Ended December 31, 2021', 2021; Sam Schechner, 'Meta's Targeted Ad Model Faces Restrictions in Europe - WSJ', 6 December 2022, https://www.wsj.com/articles/metas-targeted-ad-model-faces-restrictions-in-europe-11670335772.

Google has urged "new forms of oversight" that "address systemic, recurring failures, not one-offs", adding that "the scope and complexity of modern platforms requires an approach that focuses on overall results rather than anecdotes".[25]

Yet, there are also concerns that ramping up governmental-linked regulation is based on a misconception that official controls can be a "magic bullet" to end frustrations, and further that an overfocus on this kind of intervention sidelines the range of wider possibilities and actors in a mix of regulatory arrangements. There are also well-founded fears that increasing direct legal controls over the platform companies (where feasible) carries risks of political capture of their services. An illustration is where regulatory authorities, through steps that fall below international standards of independence and impartiality, act to shape content for narrowly political reasons. Another problem is when inappropriate laws on content allow for selective persecution against dissidents' democratic right to freedom of expression, while meanwhile also leaving intact the status quo on the platforms. Since 2020, the Global Network Initiative has identified concerns with regulatory measures for platforms in at least 22 countries.[26]

A study of 11 African countries for the period 2016 - 2020 found a doubling of laws penalising content declared "false" by authorities from 17 to 31 instances, and mainly used against critics of government.[27] Relatedly, UNESCO's own report on World Trends in Freedom of Expression and Media Development found that, between 2015 and 2020, 85% of the world's population experienced a decline in press freedom with new laws and policies restricting freedom of expression online.[28] (Section V below further unpacks legislative challenges).

It is against this backdrop that various stakeholders, and particularly human rights defenders operating in authoritarian circumstances, express caution against new restrictions that create new content crimes (like 'fake news'). They propose instead that legal regulation of platforms in regard to content should focus on deeper issue of standards and processes underlying the companies' own content controls, rather than the more surface level which where specific areas or cases of potentially problematic content are manifest. In this approach, regulation should cover matters such as transparency around platforms' business models and ad-tech, along with requirements for greater disclosure on policies and practices in content moderation and algorithmic curation. This kind of statutory regulation would avoid short-term and arbitrary approaches to the myriad items of content online, and provide instead for systemic, structural and process-based solutions.

Amongst such solutions would also be rules that require platforms to do pro-active and timely due diligence on the upholding of human rights in the face of current and foreseeable challenges. The rules should also aim to improve impact assessment efforts, as well as provide human rights compliant redress mechanisms and benchmarks for responsiveness. Taking greater account of multilingualism through regulations to incentivise and subsidise local content production and local platforms is another area that can be addressed.

Such a broader legal regulatory framework can also set out conditions for recognising industry self-regulatory structures and mechanisms, while also respecting platform pluralism as well as institutionalising multi-stakeholder participation in content policy and practice. The advice of the Working Group under the Paris Peace Forum is that the "governance process for qualifying 'harmful' content should be based on a broad and inclusive mode of participation among all stakeholders – public authorities, enterprises, civil society - that can best represent divergent interests and values and make the issue of content regulation subject to a genuine public debate".[29]

The advice of the Working Group under the Paris Peace Forum is that the "governance process for qualifying 'harmful' content should be based on a broad and inclusive mode of participation among all stakeholders.

25 Google YouTube, 'Information Quality & Contenet Moderation', n.d.

26 GNI, 'Request for Proposals: Research on Emerging Regulation on Corporate Human Rights Risk Assessment/Due Diligence', Global Network Initiative, 2022, https://globalnetworkinitiative.org/request-for-proposals-corporate-human-rights-risk-assessment/.

27 Peter Cunliffe-Jones et al., *Misinformation Policy in Sub-Saharan Africa*, *University of Westminster Press* (University of Westminster Press, 2021), https://doi.org/10.16997/book53.

28 UNESCO, 'Journalism Is a Public Good: World Trends in Freedom of Expression and Media Development; Global Report 2021/2022 - UNESCO Digital Library', 2022, https://unesdoc.unesco.org/ark:/48223/pf0000380618?2=null&queryId=0a30ee11-7640-48c0-b1c3-8d7e1e5dc867.

29 Harmful Content Working Group, 'Progress Report WHITE PAPER | Paris Peace Forum, November 2022', 2022, https://parispeaceforum.org/wp-content/uploads/2022/11/20221105_Harmful-Content-Working-Group-Report_FINAL_MIS-EN-PAGE_v2.pdf

In all this, in order to safeguard the protection of human rights online, laws should ensure that official regulatory institutions, although constituting part of the executive state apparatus, should be wholly independent of the government of the day and be primarily accountable to legislatures for fulfilment of their mandates. Where governments appoint members of regulatory instances who can instruct a platform to take down certain content, this does not meet the criteria for independence.[30]

The points above demonstrate that governmental attempts to regulate are not a panacea for the problems, and may in cases worsen it. It is also worth noting the advice of the Harmful Content Working Group under the Paris Peace Forum.[31] This observes that "public authorities do not have the material means, access and skills to police thoroughly the content shared on platforms", and so in the end "private companies - and ultimately their front-line moderators and software - are the crucial enforcement agents."

"Content moderation isn't just about individual pieces of content. It's often about patterns of behaviour, and examining how an account behaves – and what is the best intervention among a range of actions that could be taken." Marietje Schaake and Rob Reich[32]

Internationally, there are established norms for regulatory options for private publishers, private broadcasters and private telecoms operators. But the platform companies, which deal mainly with third-party content, are not directly equivalent to any of these entities which is partly why the scope of their governance regime has largely been left to be internally decided within each enterprise.

While serving as digital intermediaries, the platforms are nevertheless neither content neutral, nor are they mere conduits for the interactions of external parties. They choose their business models and they decide on terms of use for their services, both of which have major implications for content. Though the rules on their domains are often named as "community" standards, these emanate from the company concerned rather than from any actual community of users. Further, in the case of Meta and Twitter, it is the particular individuals who control the company who make the calls, effectively with accountability only to themselves.[33]

At base, private platforms typically operate policies and practices to avoid legal implication in copyrighted content or child sexual abuse materials. There is usually much else besides in their policies, covering what communications may and may not occur on their services. Not all content that is legal in a given jurisdiction is necessarily allowed on a given platform. Policies are therefore often narrower than legally required. These policies include provisions on sharing methods and types of content, and the extent of "walled-garden" restraints, application programme interfaces (APIs) and interoperability potential. There are sometimes even parameters about hyperlinks to sites outside the platform concerned[34]. Also present in platforms' policies typically are other conditions that affect content – such as provisions about who their users may be (eg. above a certain age, have a 'verified' identity or not) and how these users should behave (eg. if they may share content without the 'friction' of being encouraged to read it first). Some companies also have policies on paid-for content, such as what may be advertised and traded, (and about disclosure or not of sponsorships behind "influencers" which is significant in the context of one quarter of internet users being estimated to watch influencer videos each week[35]). Platforms enforce the range of these policies across the scale of their services to greater or lesser degrees, but generally this is with limited accountability to those who use them or to external actors.

30 Al Jazeera, 'New Delhi gives itself power over social media content moderation', 2022, https://www.aljazeera.com/economy/2022/10/28/in-india-govt-now-has-power-over-social-media-content-moderation

31 Harmful Content Working Group, 'Progress Report WHITE PAPER | Paris Peace Forum, November 2022'

32 Marietje Schaake and Rob Reich, 'Election 2020: Content Moderation and Accountability' (Human Centered Artificial Intelligence, 2020), https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/hai_cyberpolicy_election_3_v1.pdf.

33 For example, see Ryan Mac and Craig Silverman, 'Mark Changed The Rules": How Facebook Went Easy On Alex Jones And Other Right-Wing Figures', February 22, 2021, https://buzzfeed.com/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones; John Bowden, 'Free speech 'absolutist' Elon Musk personally ordered the Twitter suspension of left-wing activist, report claims', 2023. https://www.independent.co.uk/news/world/americas/us-politics/elon-musk-chad-loder-twitter-b2271556.html

34 For example, Google YouTube, 'External Links Policy - YouTube Help', n.d., https://support.google.com/youtube/answer/9054257?hl=en&ref_topic=9282365.

35 Kemp, 'Digital 2022'.

Much as a publishing enterprise can operate its own editorial policy within legal boundaries, it is broadly accepted that it is the right of private platforms to set their own service parameters within local law. In both cases, both sets of enterprises nevertheless have international obligations to still do their best to respect international human rights standards when local laws or the enforcement thereof breaches these benchmarks.

In ways similar to the distinctiveness of publishers' right to editorial autonomy, platforms are sometimes differentiated from actors acting mainly at other "layers" of the "tech stack" who operate under a general (but varying) ethos of "common carrier" [36] and/or "net neutrality" (in one of several interpretations thereof). This means they ought not to discriminate at the content or user level. Several cases do exist where non-platform intermediaries (eg. in internet service provision, digital security, and cloud services) have engaged in degrees of content-related gatekeeping – not always based on clear policy or terms of service. At the same time, infrastructure-level moderation is usually less about individual content items, but rather constitutes a sort of meta-moderation (for example about child sexual abuse materials).[37] Acknowledging therefore the different character of each layer of the Internet,[38] the content issue is more acute for platforms whose public interface operates on the top level of the "tech stack". The platforms in this role constitute a "last" stage of content moderation in the intermediation process in that they operate user-to-user and advertiser-to-user interfaces. Their functionalities accordingly range across social networking, messaging, search, advertising, and provision of marketplaces for on- or offline transactions, with the associated content and communications dimensions that go with these.

Several cases do exist where non-platform intermediaries (eg. in internet service provision, digital security, and cloud services) have engaged in degrees of content-related gatekeeping – not always based on clear policy or terms of service.

In the sense set out here, platforms therefore serve (at least in part) to intermediate content (including advertising), raising many questions about how they do so, and one question is how their policies and practices relate to general consumer protection standards (including their respect for terms of service and their offer of effective redress). The main contemporary concern goes beyond this, however, and are about how platforms deal with content that attacks other users' rights to exercise freedom of expression (for example, through threats and intimidation) and how they respond to expression that violates rights to safety, dignity, reputation, rights of the child, etc. These challenges have societal significance and point to the matter of of public accountability of platforms. In turn, this sets the stage for guidance about optimum regulatory modalities that would better protect human rights online, including freedom of expression and access to information.

All these elements are relevant to the overarching question of how governance and regulatory possibilities could lead to changes in the platforms' significance for "information as a public good" – at least inasmuch as leading to improved combating of disinformation and hate speech that run counter to this value.

36 See inter alia Reid, Blake Ellis, 'Uncommon Carriage', University of Colorado Law Legal Studies Research Paper No. 22-20, 2023, https://ssrn.com/abstract=4181948 or http://dx.doi.org/10.2139/ssrn.4181948

37 Christoph Busch, 'SPECIAL ISSUE: GOVERNING THE DIGITAL SPACE', *UCLA Journal of Law and Technology*, Governing the Digital Space, 27, no. 2 (2022).

38 Joshua Gacutan and Niloufer Selvadurai, 'The Relevance of Internet Architecture to Law: The Liability of Internet Service Providers for Harmful User-Generated Content', *ANU Journal of Law and Technology* 3, no. 1 (29 June 2022): 55–73.

3 Additional insights for regulating platforms

Understanding the term "regulation" as the design and application of rules (which can be via a variety of actors and arrangements), is best done by contextualising it in the wider ecosystem of digitally-related governance. As formulated by the World Summit on the Information Society, "Internet governance is the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet".[39]

This conceptualisation recognises a continuous spectrum ranging from principles through to programmes, and also allows that each stakeholder sector has a contribution to play at each point. At an international level, norms are shaped by initiatives such as the Internet Governance Forum, the Freedom Online Coalition, and the Declaration for the Future of the Internet. UNESCO's consultations around guidance for platform regulation are a significant contribution to the principles realm. Currently platforms operate with rules, decision-making procedures and programmes that are generally exclusive to themselves and with concomitant problems as shown in Parts 1 and 2 of this series. The question this invites is how the involvement of other constituencies into these areas of regulation can improve performance. Legislatures and executives are obvious actors to shape this realm by means of legal measures, as is their role, but they are not the entirety of regulatory possibilities.

Significant to a wider perspective on regulatory arrangements is the understanding of governance as broader than the role of governments, in the formulation of UNESCO's concept of Internet Universality. This concept advocates for not just the participation of multiple stakeholders in the various areas of governance, but also for co-operation amongst these stakeholders as well.[40] The value of all this is that actors interested in regulating platforms can take cognisance in particular of the contribution to governance of news media, regulators' associations, NGOs, platform employee bodies, whistleblowers and researchers. Their roles contribute to norms and principles, but can also be directly impactful on platforms' own rule-making and implementation.

39 Working Group on Internet Governance. 2005. "Report of the Working Group on Internet Governance." http://www.wgig.org/WGIG-Report.htm

40 UNESCO, "Background", https://www.unesco.org/en/internet-universality-indicators/background; Anri van der Spuy, 'What if we all governed the internet. Advancing multistakeholder participation in Internet governance, 2017. https://en.unesco.org/sites/default/files/what_if_we_all_governed_internet_en.pdf

MEDIA'S CONTRIBUTION TO THE GOVERNANCE ECOSYSTEM

- From troves of news reports, come the insight that independent journalists are especially key for monitoring platforms' performance, as well as for producing verified news in the face of disinformation on the platforms. Meta filings with the US Securities and Exchange Commission show high sensitivity to unfavourable media reports.
- The literature shows, however, that most news media institutions are weakened by a loss of advertising to the data-loaded digital intermediaries. In addition, researchers Asa Royal and Philip M. Napoli assess that "Platforms have also tinted the windows of story discovery, guiding users' access to news with algorithmic content curation systems that favor emotionally charged and engagement-inducing content, veracity not necessarily withstanding." Against this backdrop, there is growing momentum to mandate that platforms negotiate payments to news media. Media, however, continue to lack access to platform data that is key to their economic viability, especially that regarding the spread of news and the value of related advertising via the platforms.
- Increasingly being documented is systemic abuse of online services to attack journalists, and of this happening with relative impunity from those who control the platforms. An example of abuse is reported in a human rights impact assessment of Meta's platforms in the Philippines, which showed that a third of 75 surveyed journalists had received online death threats. Meta's corporate human rights policy specifically refers to journalists within the frame of human rights defenders, and the company offers digital security and safety training. UNESCO's The Chilling research project, conducted by International Center for Journalists (ICFJ), has extensive data about online violence against women journalists, as well as a critical assessment of platforms' responses. However, it confirms that journalists want the platforms to do much more to protect them and to proscribe their online assailants.

Regulatory guidance can encompass a role for state-based policy and law to recognise and support the place of independent journalism in the information ecosystem, as a major factor for information as a public good and for accountability of platforms to better address potentially harmful content.

Other actors relevant to platform governance:

Worldwide, national regulatory environments involve very different official regulators (and national configurations thereof), with very varied degrees of independence and a host of different national mandates under which they work. Nevertheless, initiatives like the "Global Online Safety Regulators Network"[41] and various regional networks aim to be effective in sharing experiences, for example regarding codes of conduct to serve as standards for platforms' policies and practice on a range of content matters.

It is very evident that civil society activism results in numerous investigations which put pressure on both platforms and governments to address content of potential harm to human rights.[42] Additionally, there are efforts led by civil society to professionalise employees working in "trust and safety" and to raise ethical awareness amongst those developing AI.[43] Engineers working on safety-tech in the UK have set up the Online Safety Tech Association.[44] The emergence of whistleblowers is another positive factor for platform governance that protects human rights. Coalitions for monitoring platform performance in elections[45], and networks for fact-checking have taken up some of the challenges of platform shortfalls and regulatory gaps. These can all be recognised as indispensable elements in the governance ecosystem, influencing the rules, decision-making, etc. at both government and platform level.[46]

41 E-Safety Commissioner, Australia, 'The Global Online Safety Regulators Network', 2022 https://www.esafety.gov.au/about-us/who-we-are/international-engagement/global-online-safety-regulators-network

42 See for example the Global Platform AccountabilityMovement (Ucciferri, L., nd)

43 For example, the Trust and Safety Professional Association, https://www.tspa.org/; Integrity Institute , https://integrityinstitute.org/ ; Everything in Moderation, https://www.everythinginmoderation.co/

44 https://ostia.org.uk/

45 For example, Election Integrity Partnership, https://www.eipartnership.net/; International Fact Checking Network, https://www.poynter.org/ifcn/

46 Guillaume Caline, 'Présidentielle 2022 : comment mesurer l'impact des « fake news » sur les électeurs ?', Fondation Jean-Jaurès, accessed 21 December 2022, https://www.jean-jaures.org/publication/presidentielle-2022-comment-mesurer-limpact-des-fake-news-sur-les-electeurs/; Krimmer et al., 'Elections in Digital Times: A Guide for Electoral Practitioners - UNESCO Digital Library'.

ACTIVISTS' INSIGHTS

Thanks to a number of whistleblowers within the platforms, there is more public knowledge about the problems within these companies. Research for UNESCO, conducted by the Signals Network, yielded the following insights from 11 key informants:

- Content moderation is a by-product of larger, systemic issues such as the structure of the platform and its business model.
- Regulation should mirror consumer-safety regulation in other industries must remove the financial incentive for digital content platforms to amplify harmful content.
- Democratising transparency is a fundamental need
- Support is needed for whistleblowers, investigative journalists and content moderators.

Academic researchers could be a direct and indirect asset when it comes to platform governance, but at present they largely have little option but to examine the platforms from the outside, and their insights are constrained by a lack of access to platform data. Without legal change, this situation appears to be worsening.[47] Nevertheless, there are steps to try and reverse the trend through a process engaging academics, platforms and the EU.[48] Another initiative in this area is to develop a modular approach that could be of broader relevance for international co-operation.[49]

47 Davey Alba, 'Meta Pulls Support for Tool Used to Keep Misinformation in Check - Bloomberg', Bloomberg, 2022, https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check?leadSource=uverify%20wall; Michael Grass, '
Twitter's API access changes could mark 'end of an era' in academic research on the platform, 2023 https://www.cip.uw.edu/2023/02/02/twitters-api-access-changes-academic-research/.

48 EDMO and Institute for Data Democtracy and Politics, 'Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access'; IRIE, 'About - IRIE', IRIE, n.d., https://informationenvironment.org/about.

49 Riley, 'A Module Playbook for Platform-to-Researcher Data Access'.

4 The challenging role of statutory regulations.

A large number of countries have introduced special rules against online disinformation[50], as well as sought to apply offline hate speech restrictions to the platforms. According to one study, at least 20% of 100 cases of recent laws on disinformation are related to election issues.[51] Courts too have taken highly-consequential judicial decisions in relation to content on the platforms,[52] including in relation to laws proposed in the USA's states of Florida and Texas.[53]

Executive interventions have included largely-disproportionate Internet and social media shutdowns.[54] Actions by governments calling on companies to remove content are evident in voluntary reports by some platforms, although further granular details about the official rationales and the platforms' responses are often not generally supplied.[55] Local law has seen some companies reporting on relevant content restrictions they have imposed.[56] However, a lacuna is that authorities are often not required to report publicly on their demands to companies. There is also little evidence of officials being prosecuted for fomenting online hate and disinformation under the relevant regulations.[57]

Various other challenges have been flagged:

In many countries, there is public distrust of governmental efforts to regulate online content.[58] Given the widespread role of political forces in generating disinformation, as discussed in Part 1 of this series, there may be a conflict of interests regarding their involvement in regulating platforms. In addition, many regulators are far from being independent.[59] Further, the capacity of many smaller states to apply legal measures on international platforms is limited, which runs against the core principle of regulation which is to avoid issuing prescriptions that cannot be enforced. Enforcement in turn depends on capacity to monitor and assess, which is also a challenge for all states since it is compounded by platform opacity and complexity.

50 Krimmer et al., 'Elections in Digital Times: A Guide for Electoral Practitioners - UNESCO Digital Library'; Par Dorian Mouketou, 'École nationale d'administration publique (ENAP)' (École nationale D'administration Publique, 2021), https://espace.enap.ca/id/eprint/250/1/Mouketou,%20Dorian_STA_210604.pdf.

51 Kamya Yadav, 'Platform Interventions: How Social Media Counters Influence Operations - Carnegie Endowment for International Peace', Carneige Endowment for International Peace, 2021, https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698.

52 Irène Couzigou, 'The French Legislation Against Digital Information Manipulation in Electoral Campaigns: A Scope Limited by Freedom of Expression', *Election Law Journal: Rules, Politics, and Policy* 20, no. 1 (1 March 2021): 98–115, https://doi.org/10.1089/elj.2021.0001.

53 Ramya Krishnan, 'How the Supreme Court Could Encourage Platform Transparency without chilling free speech', 2023, https://slate.com/technology/2023/01/supreme-court-florida-texas-social-media-laws.html

54 Lisa Garbe, Lisa-Marie Selvik, and Pauline Lemaire, 'How African Countries Respond to Fake News and Hate Speech', *Information, Communication & Society*, 9 November 2021, 1–18, https://doi.org/10.1080/1369118X.2021.1994623; Bosompem Boateng Richard, *Social Media Usage and Digital Rights Restrictions in the Republic of Chad, Digital Dissidence and Social Media Censorship in Africa* (Routledge, 2022), https://doi.org/10.4324/9781003276326-14.

55 TikTok, 'Government Removal Requests Report Jul - Dec 2021 | TikTok', 2021, https://www.tiktok.com/transparency/en-us/government-removal-requests-2021-2/; Twitter, 'Removal Requests - Twitter Transparency Center'. https://transparency.twitter.com/en/reports/information-requests.html#2021-jul-dec

56 Meta, 'Content Restrictions Based on Local Law | Transparency Center' (Meta, 2022), https://transparency.fb.com/data/content-restrictions/.

57 Cunliffe-Jones et al., *Misinformation Policy in Sub-Saharan Africa*.

58 Shilongo Kristophina, Hanani Hlomani, and Araba Sey, 'Responses to Information Disorders: What Can Governments Do?', 2022, https://www.africaportal.org/publications/responses-information-disorders-what-can-governments-do/.

59 Gillespie and Aufderheide, 'Expanding the Debate About Content Moderation'.

Research reveals problems where local definitions of content that is deemed to be potentially harmful are too vague or wide to meet the test of predictability, and that also allow for politically selective application.[60] The same applies to legal definitions of what constitutes a "platform"[61] and "amplification"[62]. It has been pointed out that using the term "fake news" in law provides much leeway in deciding what content is problematic or false, and the indirect possibility that political influence will also define what content is true[63].

Hate speech is criminalised under international standards when it counts as advocacy to incitement for violence, hostility or discrimination, aimed in this case against categories of persons. But disinformation and misinformation are not illegitimate unless harnessed for such purposes, or for other rights-violating purposes such as health, safety, dignity and reputation.[64] However, laws and other statutory regulatory initiatives do not always make these distinctions. Instead, they often enable blanket criminalisation of all content deemed by authorities to be expression of falsehoods or revulsion, even if the utterances concerned pose no threat to human rights.

Penalties are frequently disproportionate to defined "content crimes". An analysis of 10 of 31 existing and new laws in sub-Saharan Africa showed there was no evidence required that actual harm resulted from the content at stake.[65]

In many cases, platforms are legally protected through conditional liability for content carried, with the liability placed exclusively for those posting the content in the first place. Where such content is illegal, liability applies only if companies are self-aware of, or been notified by a third party, of such content and left it online. The overall lack of legal obligation for the platforms is widely seen as responsible for an insufficient fulfilment of duties of care in order to protect human rights on their services.

In many cases, platforms are legally protected through conditional liability for content carried, with the liability placed exclusively for those posting the content in the first place.

60 Mark Verstraete, Derek E. Bambauer, and Jane R. Bambauer, 'Identifying and Countering Fake News', *SSRN Electronic Journal*, 2017, https://doi.org/10.2139/ssrn.3007971; Peter Cunliffe-Jones, 'OP-ED: Misinformation Literacy, Not Punitive Laws, Needed to Combat Fake News', Daily Maverick, 29 April 2021, https://www.dailymaverick.co.za/article/2021-04-29-misinformation-literacy-not-punitive-laws-needed-to-combat-fake-news/; Ferdaouis Bagga, 'Apostasy, Blasphemy, and Hate Speech Laws in Africa' (United States Commission on International Religious Freedom, 2019); Edzodzi Kokou Ahiadou, 'How States in Francophone West Africa Are Weaponizing Legislation to Suppress Freedom of Expression Online', Media Foundation For West Africa, 13 December 2021, https://www.mfwa.org/how-state-in-francophone-west-africa-are-weaponizing-legislation-to-suppress-freedom-of-expression-online/; Garbe, Selvik, and Lemaire, 'How African Countries Respond to Fake News and Hate Speech'.

61 Rajendra-Nicolucci and Zuckerman, *An Illustrated Field Guide to Social Media*.

62 Daphne Keller, 'Amplification and Its Discontents' (Knight First Amendment Institute at Columbia University, 2021), http://knightcolumbia.org/content/amplification-and-its-discontents.

63 Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches'.

64 Irene Kahn, Un Secretary-General, and UN Human Rights Council Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'Disinformation and Freedom of Opinion and Expression during Armed Conflicts', 12 August 2022, https://digitallibrary.un.org/record/3987899.

65 Cunliffe-Jones, 'OP-ED'.

THREE RISKS TO THE RIGHT TO FREE EXPRESSION

PRIVATISED CENSORSHIP

In some cases, there are governmental initiatives to compel companies to act against speech that is "lawful, but awful".[66] This, along with allowances for non-judicial actors to demand action from the companies, is problematic in terms of international standards on freedom of expression which requires, at least for states, that restrictions on rights be set in law and subject to arbitration by an independent judiciary. Draft laws have also been criticised for loose definitions of what constitutes "legal but harmful" content, for having overbroad scope, and for allowing legal overreach that encroaches on civil liberties.[67]

PRIOR CENSORSHIP

Companies like to publicise how content apparently violating their terms of service is intercepted through AI at the upload stage. However, there is concern that such blanket technical measures may entail over-monitoring in ways that violate privacy, as well as constitute a disproportionate type of prior restraint. Governments are advised by civil society to refrain from legally requiring this kind of action on the basis that it would equate to the state authorising a form of prior censorship. Less controversial is voluntary implementation by companies of prior restraint in narrow remits like high-priority illegal content on their platforms, such as intercepting uploads of child sexual abuse materials or halting live video streams of terror attacks. However, the public at large has little awareness of the risks that upload filtering technology can pose even in regard to publication of expression, even that which falls within platform terms of service, as well as to the public's right to know.[68]

BLANKET CENSORSHIP

International standards do exist for nuanced assessment of restricting expression, as in the Rabat Plan of Action against hate speech.[69] Such approaches could give guidance to judicial, regulatory and platform actors on dealing with cases such as when otherwise lawful expression is amplified online, linked to illegal expression and/or serving as dog whistles for violence, to the extent that human rights are seriously endangered.[70] On such a basis, platforms could be required to act pre-emptively (such as serving warnings to perpetrators to desist) in order to avoid situations where it is too late to moderate after something goes wrong.[71] At the same time, legal requirements requiring advance risk assessments could ensure that platforms set thresholds for severe danger to human rights. This that would then help to avoid unwarranted restriction in the absence of given content narratives beginning to scale and combine to constitute a credible likelihood of harm.[72]

66 Thomas Seal, 'UK Rewrites Online Safety Bill After Free Speech Backlash', 2022, https://www.bloomberg.com/news/articles/2022-11-28/uk-rewrites-online-safety-bill-to-address-free-speech-backlash?leadSource=uverify%20wall

67 Kir Nuthi and Mella Tesfazgi, 'Reforming the UK Online Safety Bill to Protect Legal Free Expression and Anonymity' (Center for Data Innovation, 2022); Adler and Thakur, 'A Lie Can Travel'.

68 Amélie Pia Heldt, 'Upload-Filters: Bypassing Classical Concepts of Censorship?', JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law, 10 (1) 2019, 56 https://www.jipitec.eu/issues/jipitec-10-1-2019/4877
Emma J Llanso, 'No amount of "AI" in content moderation will solve filtering's prior-restraint problem', Big Data & Society, January–June: 1–6, 2020, https://journals.sagepub.com/doi/pdf/10.1177/2053951720920686

69 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

70 Keller, 'Amplification and Its Discontents'

71 Joan Donovan, 'Social-Media Companies Must Flatten the Curve of Misinformation', *Nature*, 14 April 2020, d41586-020-01107-z, https://doi.org/10.1038/d41586-020-01107-z.

72 Susan Benesch, 'The Insidious Creep of Violent Rhetoric', *Digital Society*, 4 March 2021, https://www.noemamag.com/the-insidious-creep-of-violent-rhetoric.

5 A range of regulatory arrangements comprising an overall hybrid system

Of global relevance is the UN's "Our Common Agenda" which envisages a voluntary code of conduct on public information on the platforms, which will complement agreements anticipated under the Global Digital Compact anticipated in 2024.[73] This development will be a further element in the overall platform governance ecosystem.

At regional level, the EU has been a pioneer in developing, with participation with the companies, what will become stronger codes of conduct and practice for disinformation and for hate speech.[74] The initial phase of these codes was criticised for insufficient legal standing, and for inadequate monitoring, sanctioning and transparency. Several of these points will be now partly addressed under the new DSA.[75]

At national level, co-regulatory codes of conduct have been a feature in Australia. There, the eSafety Commissioner has advised that such codes should have minimum compliance measures for platforms. These include: age-'gating' through verification or age assurance mechanisms; interstitial notices; warning labels, warning/notice screens; downlisting or deprioritising content; quarantining; and image/text/audio masking. In addition, measures include reducing promotion and reach within algorithmic systems, including recommendation algorithms and choice architecture; and internal policies and procedures to proactively monitor, assess, investigate, and audit content within algorithmic systems.

Some observers warn that co-regulatory arrangements in general can blur into privatisation of censorship whereby platforms are leaned upon by authorities lacking explicit legal mandate, and where the casualty is platforms taking actions against lawful expression. Another fear is that platforms begin to collude as monolithic content cartels that reduce pluralism of content.[76] Further, age-'gating' as a requirement in a code of conduct can raise thorny questions of bulk surveillance and privacy intrusion[77]. In the absence of safeguards and strong transparency mechanisms for co-regulatory arrangements, platforms may simply decide to over-compensate in regard to imposing restrictions on content. The danger in this, to the rights to freedom of expression and political participation in particular, is that platforms ultimately become more responsive to pressure to restrict content which reflects poorly on the authorities, but which is nevertheless legitimate in terms of their terms of service, local laws and human rights norms and standards.

73 United Nations, 'Our Common Agenda', United Nations (United Nations, 2021), https://www.un.org/en/common-agenda.

74 Dawid Aristotelis Fusiek, Angeliki Elli Stougiannou, and Theoharris William Efthymiou-Egleton, 'Digital Democracy and Disinformation: The European Approach to Disinformation on Social Media in the Case of 2019 European Parliament Elections', *Journal of Politics and Ethics in New Technologies and AI* 1, no. 1 (29 August 2022): e31215, https://doi.org/10.12681/jpentai.31215.

75 Ethan Shattock, 'Self-Regulation 2:0? A Critical Reflection of the European Fight against Disinformation', *Harvard Kennedy School Misinformation Review*, 31 May 2021, https://doi.org/10.37016/mr-2020-73; European Commission, 'Guidance to Strengthen Code of Practice on Disinformation', European Comission, 2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_2585.

76 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; Robert Gorwa, 'The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content', *Internet Policy Review* 8, no. 2 (30 June 2019), https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content; Michael Karanicolas, 'Authoritarianism as a Service: India's Moves to Weaponize Private Sector Content Moderation with the 2021 Information Technology Rules', *SSRN Electronic Journal*, 2022, https://doi.org/10.2139/ssrn.4145058.

77 Rebecca Mackinnon and Phil Bradley-Schmieg, 'UK Threatens Blowtorching Internet Platforms – Including Wikipedia', 2022, https://cepa.org/article/uk-threatens-internet-platforms/?ref=everything-in-moderation

CO-OPERATIVE REGULATORY ARRANGEMENTS IN REGARD TO ELECTIONS

One approach to regulation is to "modularize" the elaboration of different areas as appropriate to the issues and prioritization at stake. [78] A modular approach to unpacking different issues in platform governance (eg. hate speech, misinformation, bullying, researcher access, risk assessment templates, etc.) is a way to tackle both institutional and functional specific issues (eg. search engines like Google, search functionalities across several companies). It can also address issues like incitement to violence that cut-across different companies and services.

In this regard, elections would seem to be a primary candidate for a modular approach. An election represents the primary mechanism for democracy and the exercise of the right to political participation. In this context, the governance of platforms' roles during polls can affect the integrity and credibility of the exercise.

Some of the challenges can be seen from UNESCO's aggregation of available data on six sample countries[79] from around the world that will hold elections in 2023 and 2024. In three of the countries combined, there are more than 100 different languages spoken. Across all six countries, the Disinfodex reports that there was already a total of 61 recorded disinformation campaigns between 2017 and April 2021. In five of these countries, the Edelman Trust Barometer finds levels of public distrust are registered above 50%. Regarding the latter, researchers William T Adler and Dhanaraj Thakur have noted in general that "(l)ow levels of trust in democracy and in government can create a vicious cycle when combined with election disinformation. For example, low trust may increase receptiveness to election disinformation, which in turn may further reduce trust in democracy."[80]

Platform performance in elections is partly a function of the specific national regulatory frameworks around such events, but many of these frameworks are outdated and weak in regard to the platforms.[81] In the absence of applicable legal provisions, the detail about platforms' treatment of online political advertising and sponsorships in particular remains largely opaque. Meta has even taken legal action to prevent researchers from independently monitoring the extent of micro-targeted political advertising on Facebook.[82]

However, even without legal compulsion, the companies for their part are increasingly proactive[83], producing policies about how they hope to deal with election issues. UNESCO analysis using a database of policy announcements related to elections, between 2020 and November 2022, shows almost 180 statements from Meta, Google, Parler, Linkedin, Apple, Hulu, YouTube, TikTok, Spotify, Amazon, Lyft, Uber, Salesforce and others.[84] Many of these policies, however, lack provisions on voter suppression and on fact-checking electoral adverts.[85]

The articulation of generic platform policies to specific national election rules and institutions is a major question, as is the record of actual implementation and review. A study of TikTok showed the company was "woefully unprepared to combat electoral misinformation" in the US.[86] In Kenya, the Mozilla Foundation found that Facebook and Instagram ran adverts violating election law, while Meta, Twitter and TikTok had failed to moderate harmful posts.[87]

In this context, guidance for regulation can promote a holistic approach covering legal steps that underline the application of national electoral specifics, provide for standards for online political advertising, and embrace formal and informal multi-stakeholder arrangements. All this could help ensure that platforms' policies or actions supporting voter information efforts, while welcome, are not a substitute for them supporting national rules for election integrity.[88] Electoral management bodies could develop also consider novel regulatory mechanisms involving multi-stakeholder participation, such as the South African experience titled "Real411" entailing a partnership against online disinformation between the country's electoral regulator, several key platforms, an NGO, editors and lawyers.[89]

---

78 Chris Riley, 'A Module Playbook for Platform-to-Researcher Data Access', Tech Policy Press, 20 November 2022, https://techpolicy.press/a-module-playbook-for-platform-to-researcher-data-access/.

79 The countries are Argentina, Australia, India, Indonesia, Mexico and Thailand.

80 Adler and Thakur, 'A Lie Can Travel'.

81 Alexandra Brown and Lisa Reppell, 'Lessons for Regulating Campaigning on Social Media', 2021.

82 Tromble, 'Facebook, NYU, and the "Risks" of Public Interest Research'.

83 Chris Vallance, 'TikTok to Teach Influencers about US Mid-Term Election Rules', *BBC News*, 17 August 2022, sec. Technology, https://www.bbc.com/news/technology-62552702; Tom Gerken, 'Google to Run Ads Educating Users about Fake News', *BBC News*, 24 August 2022, sec. Technology, https://www.bbc.com/news/technology-62644550; Meta, 'How Meta Is Planning for the 2022 US Midterms', *Meta* (blog), 16 August 2022, https://about.fb.com/news/2022/08/meta-plans-for-2022-us-midterms/; Twitter, 'Bringing More Reliable Context to Conversations on Twitter', Twitter Blog, 2021, https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter; Twitter, 'Our Approach to the 2022 US Midterms', Twitter Blog, 2022, https://blog.twitter.com/en_us/topics/company/2022/-our-approach-to-the-2022-us-midterms.{\\i{}BBC News}, 17 August 2022, sec. Technology, https://www.bbc.com/news/technology-62552702; Tom Gerken, \\uc0\\u8216{}Google to Run Ads Educating Users about Fake News\\uc0\\u8217{}, {\\i{}BBC News}, 24 August 2022, sec. Technology, https://www.bbc.com/news/technology-62644550; Meta, \\uc0\\u8216{}How Meta Is Planning for the 2022 US Midterms\\uc0\\u8217{}, {\\i{}Meta} (blog

84 See https://docs.google.com/spreadsheets/d/1qi4pmAzPXEYHxnK8Snf5KJc8T0gq4bQrcQshCds65X8/edit#gid=820055922

85 Sinders, 'The Use of Mis- and Disinformation in Online Harassment Campaigns'; Gerken, 'Google to Run Ads Educating Users about Fake News'; Katie Harbath, 'Tech Company 2022 Midterm Election Announcements | Bipartisan Policy Center', Bipartisan Policy Center, 2022, https://bipartisanpolicy.org/blog/tech-midterm-election-announcements/.

86 Olivia Little, 'Misinformation about the Midterm Elections Is Already Flourishing on TikTok | Media Matters for America', Media Matters for America, 2022, https://www.mediamatters.org/tiktok/misinformation-about-midterm-elections-already-flourishing-tiktok.

87 Andrew Deck, 'Facebook and Instagram ran ads violating Kenyan election law, new report reveals', 3 November 2022, https://restofworld.org/2022/facebook-instagram-ads-kenya-election/

88 Vallance, 'TikTok to Teach Influencers about US Mid-Term Election Rules'.

89 South African Electoral Commission, 'Multi-stakeholder partnership to combat disinformation in the 2021 Municipal Elections', 2021, https://www.elections.org.za/content/About-Us/News/Multi-stakeholder-partnership-to-combat-disinformation-in-the-2021-Municipal-Elections/

In most regulatory landscapes, an issue that calls out for attention is the intersection between national policies and associated regulatory instances (for example, on elections, discrimination, gender equality, children, migration or health). All of these are affected by how platform companies treat disinformation and hate speech.[90] Under the EU's DSA, national coordination mechanisms in the member countries are envisaged across different regulatory institutions relevant to potentially rights-harming online content. These bodies may include electoral management bodies, audio-visual regulators, advertising standards enforcers[91], privacy and data protection commissions, competition bodies, etc. However, most jurisdictions lack such coordinating mechanisms, and many will find it challenging to allocate resources to such. It is in this light that it is worth recalling how multi-stakeholder participation, industry self-regulatory, and co-regulatory arrangements can help share the load.

A further key issue for all regulatory arrangements to consider is the insight that technology continues to develop apace, such as AI. Attention is needed to "deep fakes"[92] in language[93], sound and imagery[94]; and there are issues in human rights protection regarding the "metaverse"[95] and web 3.0. One experiment with GPT-3 showed that it scaled fairly easily to produce disinformation including by hijacking viral hashtags[96]. Platform companies are developing policies on "manipulated" media[97], but are accused of having been slow to anticipate and pre-empt content problems.[98] However, the literature also shows that there is little legal regulatory attention to these issues, although there are distinctive and serious risks to human rights at stake.[99] Unlike health and automotive industries, most jurisdictions lack statutory regulatory requirements for platforms to practise Safety by Design[100] (or provide for user Autonomy by Design[101]).

Platform companies are developing policies on "manipulated" media, but are accused of having been slow to anticipate and pre-empt content problems.

90 Internet Freedoms, 'Kenya Government Dismisses NCIC'S Threats to Shut down Facebook', *IFreedoms Kenya* (blog), 2 August 2022, https://www.ifree.co.ke/2022/08/kenya-government-dismisses-ncics-threats-to-shut-down-facebook/; Axel Boursier, 'Les jeux de la vérité dans les tweets de haine anti-migrants', 2021, https://shs.hal.science/halshs-03204033/document.

91 For example, see Olatunji Olaigbe, 'Nigerian influencers could soon need government approval for sponsored posts', 21 December, 2022, https://restofworld.org/2022/nigerian-influencers-government-approval/

92 Emmie Hine and Luciano Floridi, 'New Deepfake Regulations in China Are a Tool for Social Stability, but at What Cost? | Nature Machine Intelligence', Nature Machine Intelligence, 2022, https://www.nature.com/articles/s42256-022-00513-4; Schaake and Reich, 'Election 2020: Content Moderation and Accountability'.

93 Cooper Raterink, 'Assessing the Risks of Language Model "Deepfakes" to Democracy', Tech Policy Press, 21 May 2021, https://techpolicy.press/assessing-the-risks-of-language-model-deepfakes-to-democracy/.

94 Britt Paris and Joan Donovan, 'Deepfakes and Cheapfakes' (Data and Society, 2019), https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf.

95 Nick Clegg, 'Making the Metaverse: What It Is, How It Will Be Built, and Why It Matters', *Medium* (blog), 20 May 2022, https://nickclegg.medium.com/making-the-metaverse-what-it-is-how-it-will-be-built-and-why-it-matters-3710f7570b04.

96 Ben Buchanan et al., 'Truth, Lies, and Automation', *Center for Security and Emerging Technology* (blog), 2021, https://cset.georgetown.edu/publication/truth-lies-and-automation/. Jeremy Kahn, 'This Article is Fake News. But It's Also The Work of AI', February 14, 2019, https://www.bloomberg.com/news/articles/2019-02-14/this-article-is-fake-news-but-it-s-also-the-work-of-ai?sref=QYWxDQ1o&leadSource=uverify%20wall; Stanford Internet Observatory, OpenAI, and Georgetown University's Center for Security and Emerging Technology, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations', 2023, https://cyber.fsi.stanford.edu/io/publication/generative-language-models-and-automated-influence-operations-emerging-threats-and

97 Twitter, 'Our Synthetic and Manipulated Media Policy | Twitter Help', 2022, https://help.twitter.com/en/rules-and-policies/manipulated-media.

98 Jamie Cohen, 'Researchers Have Warned About Harassment in the Metaverse for Decades', *OneZero* (blog), 22 February 2022, https://onezero.medium.com/researchers-have-warned-about-harassment-in-the-metaverse-for-decades-c66c6293bced.

99 Center for Countering Digital Hate, 'Malgorithm. How Instagram's Algorithm Publishes Misinformation and Hate During the Pandemic'; CCDH, 'Facebook's Metaverse', Center for Countering Digital Hate | CCDH, 2021, https://counterhate.com/research/facebooks-metaverse/; Julie Owono and Leïla Morch, 'Content Governance in the Metaverse' (Content Policy and Society Lab, 2022).

100 Australia eSafety Commissioner, ' Safety by design', https://www.esafety.gov.au/industry/safety-by-design; CCDH, 'Star Framework. A Global Standard for Regulating Social Media'.

101 Christian Djeffal et al, 'Recommender Systems and Autonomy: A Role for Regulation of Design, Rights, and Transparency'. 2022. https://www.researchgate.net/profile/Christian-Djeffal/publication/358703721_RECOMMENDER_SYSTEMS_AND_AUTONOMY_A_ROLE_FOR_REGULATION_OF_DESIGN_RIGHTS_AND_TRANSPARENCY/links/620fb11e4be28e145c9e8fca/RECOMMENDER-SYSTEMS-AND-AUTONOMY-A-ROLE-FOR-REGULATION-OF-DESIGN-RIGHTS-AND-TRANSPARENCY.pdf

Recommendations

- Regulatory arrangements in all forms and by all actors need to mainstream a human rights approach; statutory ones should be articulated with other official policy areas and regulatory bodies, and avoid the range of pitfalls which can erode freedom of expression.
- Statutory law and regulation should encompass not only what should not be done (such as platforms not amplifying illegal content), but also what should happen (for instance, greater transparency, rule-governed moderation process and independent impact assessments).
- An international guidance framework could provide a common reference point for emerging and fragmented regulatory regimes, and for supporting decentralised platform alternatives such as governmental presence on non-profit and decentralised services like the fediverse.
- Guidance can also help to unpack the value of combining various regulatory arrangements into a hybrid overall system including legal delegation of roles to platforms, mandated codes of conduct, and voluntary standards operating with regard to government-linked but independent regulation.
- Guidance can inform mechanisms for co-regulation, and the extent to which the latter is part of a statutory dispensation or more informal. Legal regulation for multistakeholder participation, such as requiring regulators and platforms to engage vulnerable groups and those victimised by intersectional characteristics, may also be of value.
- The roles of media, NGOs, tech employee bodies, whistle-blowers and researchers should be recognised as positive elements in the wider governance ecosystem in which regulatory arrangements take place.
- Following UNESCO's concept of "Internet Universality", institutionalised multi-stakeholder roles can be proposed regarding all regulatory aspects, including the generation of platform policy, monitoring, appeal systems, as well as evaluation.
- Attention can be drawn to the value of a variety of regulatory arrangements that deal with the range of platforms, issues, different regulators and other stakeholders.
- Guidance can signal the importance of anticipating technology evolutions, and require platforms to conduct relevant human rights impact assessment, practice Safety by Design and Autonomy by Design, and checks by highly trained personnel on automated decision-making.
- Guidance on implementation strategies could encourage modularity approaches, within a wider and comprehensive human rights perspective, as well as call for more transnational cooperation between stakeholders such as on regulatory modalities and experiences.

Call for input

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

- Are there inaccuracies or omissions?
- What design features for process and outcomes-based regulatory systems can better deal with content issues that new content-focused laws purport to be responding to?
- How do various regulatory arrangements deal with the challenge to identify rights-harming content without engaging in bulk monitoring which may intrude on privacy rights?
- How can an overall hybrid system best address the diversity of regulators, the range of actors in the "tech stack", and the issues of capacity and political capture that constrain effective regulation in many countries?
- In what ways might region-wide co-operation and even regulation, especially in developing countries, deal with the challenges of those platform companies offering local services but are outside national jurisdictions?
- In what ways could statutory regulation institutionalise multi-stakeholder involvement across the range of digital governance elements, especially for direct and indirect rule-making and enforcement at the level of states, the platform sector, and individual platform companies?
- Besides attention focused upon regulatory arrangements that address platforms' roles in elections, what other types of "modules" could be considered as potentially shared priorities amongst a range of countries?

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

This research was supported by UNESCO