MATTHIAS C. KETTEMANN AND WOLFGANG SCHULZ (EDS.)

Platform://Democracy

Perspectives on Platform Power, Public Values and the
Potential of Social Media Councils

PLATFORM://DEMOCRACY

Platform://Democracy

Perspectives on Platform Power, Public Values and the Potential of Social Media Councils

edited by Matthias C. Kettemann and Wolfgang Schulz (eds.)
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY,
ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT

ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY

STIFTUNG MERCATOR

Executive Summary

Ground Rules for Platform Councils – Results from the *Platform://Democracy*[1] project

Social media platforms have created private communication orders which they rule through terms of service and algorithmic moderation practices. As their impact on public communication and human rights has grown, different models to increase the role of public interests and values in the design of their rules and their practices have, too. But who should speak for both the users and the public at large? Bodies of experts and/or selected user representatives, usually called Platform Councils or Social Media Councils (SMCs), have gained attention as a potential solution.

Examples of Social Media Councils include Meta's Oversight Board but most platform companies have so far shied away from installing one. This survey of approaches to increasing the quality of platform decision-making and content governance involving more than 30 researchers from all continents brought together in regional "research clinics" makes clear that trade-offs have to be carefully balanced. The larger the council, the less effective is its decision-making, even if its legitimacy might be increased.

While there is no one-size-fits-all approach, the project demonstrates that procedures matter, that multistakeholderism is a key concept for effective Social Media Councils, and that incorporating technical expertise and promoting inclusivity are important considerations in their design.

As the Digital Services Act becomes effective in 2024, a Social Media Council for Germany's Digital Services Coordinator (overseeing platforms) can serve as test case and should be closely monitored.

Beyond national councils, there is strong case for a commission focused on ensuring human rights online can be modeled after the Venice Commission and can provide expertise and guidelines on policy questions related to platform governance, particularly those that affect public interests like special treatment for public figures, for mass media and algorithmic diversity. The commission can be staffed by a diverse set of experts from selected organizations and institutions established in the platform governance field.

Table of Contents

Research Report Europe **146**

How to Integrate Public Values into Private Orders

Results from the "Platform://Democracy"[2] project

Matthias C. Kettemann and Wolfgang Schulz[3]
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG

Summary

– **How to make digital spaces more democratic?** Social media platforms have created – in principle – legitimate private communication orders which they rule through terms of service and algorithmic moderation practices. As their impact on public communication and human rights has grown, different models to increase the role of public values in the design of their rules and their practices have, too. But who should speak for both the users and the public at large? Bodies of experts and/or selected user representatives, usually called Platform Councils of Social Media Councils (SMCs) have gained attention as a potential solution. To ensure public values in these hybrid communication spaces, private actors have been asked to legitimize their primacy and incorporate public elements, such as participatory norm-setting processes and ties to human rights. The Oversight Board by Meta represents one of the first attempts to open up the decision-making system of a commercial platform to the "outside".

– **How to design platform councils?** The idea behind platform councils is to increase inclusivity in decision-making and communication space-as-product design. However, there is limited evidence on how to best construct a platform council, and the most effective approach seems to be one of regulated self-regulation. This means combining commitments by states to develop a normative framework and engaging platforms' interest to meet compliance obligations. Platform councils could be established at different levels in different set-ups in order to iterate attempts to increase the legitimacy of platform rules and algorithmic communication governance orders.

– **How global should platform councils be?** Different types of social media councils (national, regional, and global) can be effective in addressing social media governance issues that vary in scope, context, and objectives. Multi-level collaboration and coordination among different levels of councils may be necessary to effectively address the complex and multifaceted challenges of social media governance.

– **Should they be manned by experts or as broad as possible?** Incorporating technical expertise into platform council decision-making is crucial, but requires balancing the need for pertinent information with the importance of ensuring the independence of the expert's perspective. Inclusivity and promoting marginalized and minority groups are important considerations in the design of social media councils.

– **Multistakeholderism as a key concept:** To effectively address both global and local challenges and interests, SMC governance structures must be tailored to meet the needs of all, including non-users and non-onliners. This is where the concept of multistakeholderism comes into play. SMCs should have a diverse range of members, including users, experts, and citizens, and should be set

---

up through a multi-stakeholder process that involves all relevant actors. Together with public consultation measures this ensures that the decisions made by the SMC are fair, transparent, and representative of all stakeholders' interests.

– **Drawbacks and trade-offs:** While SMCs can provide more legitimacy to the rules and algorithmic practices of platforms, there are also drawbacks and trade-offs to consider. These include a weakening of state regulators, confusion of responsibility, ethics-washing, a normative fig-leaf effect, and a globalist approach to speech rules that is non-responsive to local and regional practices. Additionally, incorporating too many non-experts into the assessment of rules and algorithmic practices can be problematic. The need for access to specific information about a platform's operations raises issues of confidentiality, which can limit the effectiveness of an SMC.

– **Europe's approach:** The new European digital law packages include references to civil society integration into the national compliance structures of the Digital Services Act without specifying how exactly this process should be implemented. The German national implementation law includes a reference to a cross-platform social media council to be set up by parliament to advise the national Digital Services Coordinator.

– **What way forward?** Beyond national councils, there is strong case for a commission focused on ensuring human rights online can be modeled after the Venice Commission and can provide expertise and guidelines on policy questions related to platform governance, particularly those that affect public interests like special treatment for public figures, for mass media and algorithmic diversity. The commission can be staffed by a diverse set of experts from selected organizations and institutions established in the platform governance field.

Increasing the justice of hybrid speech orders

– "We are not a democracy", the internal rules of AI-based image-generator, the research lab Midjourney, proclaim. "Behave respectfully or lose your rights to use the Service". Indeed, companies are not democracies, they are not run by elected representatives and their business models are not voted on directly – only indirectly through the power of users, if the markets they are active in function well.

– And yet: the spaces of communication and the communicative infrastructures of democratic public spheres are subject to considerable processes of change. As the European Court of Human Rights noted in its 2015 *Cengiz* ruling, "the Internet has now become one of the principal means by which individuals exercise their right to freedom to receive and impart information and ideas, providing [...] essential tools for participation in activities and discussions concerning political issues and issues of general interest." Basic questions of our society are being negotiated in spaces that are increasingly algorithmically optimized and normatively shaped by digital platforms.

– Socio-communicative spaces have not only been enriched by taking on a digital dimension, they have also changed in character: a majority of online communication takes place in privately owned and regulated communicative settings. The key questions regarding how to enable, moderate and regulate speech today have to be asked and answered with a view to private digital spaces. These changes in communicative spatiality take nothing away from the primary responsibility and ultimate obligation of states to protect human rights and fundamental freedoms, online just as offline. However, over the last decade, tension between the normativity inherent in the role of states and the facticity of online communicative practices that are being primarily regulated by the rules of private actors has increased in intensity.

– Traditional approaches to the involvement of citizens in the development of speech rules for online space do not work as tried and tested democratic principles cannot easily be translated to allow user participation in the design of private selection algorithms and moderation practices. Just as platforms themselves have become rule-makers, rule enforcers and judges of their decisions,

politicians have recognized the importance of considering public contributions to private speech rules. In the German ruling coalition's multi-party agreement, the government committed itself to "advancing the establishment of platform councils". In a similar vein, the German Academies of Sciences and Humanities called for the participation of "representatives of governmental and civil society bodies as well as (...) users (...) in decisions about principles and procedures of content curation".

– States – through laws and courts – have laid down ground rules for platforms to accept, for example in relation to how they treat users and how to handle illegal content. To a growing extent, therefore, private normative orders and state legal orders are interacting to create neither solely private nor public, but *hybrid* forms of governance of online communication. This makes eminent sense: the balancing of interests on platform-based online communication invokes issues of public interest, but is the result of the private *product* design of platform providers.

– Orders of power are either public or private or hybrid. Private orders – often called "Community Standards" –, based on contracts, are legitimate and often successful in regulating communication spaces. In the last years, private orders have been infused with public elements, as these communication spaces are increasingly carrying democratic discourse, thus creating a hybrid order of speech governance.

– Hybrid normative orders are characterized by both private and public features relating to ownership (often private), scale (of public impact), participants (both private and public, i.e. state, actors), and values (private goals, sometimes aligned with, but sometimes conflicting with public values). Within hybrid orders, difficult normative questions emerge as to the application of fundamental and human rights (Drittwirkung; horizontal application) and the role that third parties (non-users, the public, society) should play.

– States have started to regulate private and hybrid communication spaces with a view to fulfilling their obligation regarding coordination and cooperation to ensure public values. At the same time, private actors have been asked to further legitimize their primacy over communication spaces. This is a challenge that the founder of the Leibniz Institute for Media Research | Hans-Bredow-Institute, the pioneer of German broadcasting, Hans Bredow, wrote about 75 years ago. He suggested the creation of a council to represent the population in broadcasting decisions. The challenge now was to learn from this approach in the digital age, and especially with regard to private and hybrid communication spaces.

– In this project, when we speak of "enhanced legitimacy" of norm-setting and norm-enforcement by platforms, we mean that elements are added that make an impact of private action on public communication appear (more) just. This can happen by making the norm-setting process itself more participatory, but also by tying it more back to human rights.

On the development of platform councils

– In 2021 we published an introductory study on Social Media Councils, exploring the concept and their origins in media councils. To build trust, platforms should try a little democracy suggested Casey Newton. ARTICLE 19 published a report on their Social Media Councils experiment in Ireland. Towards Platform Democracy: Policymaking Beyond Corporate CEOs and Partisan Pressure | Belfer Center for Science and International Affairs is Aviv Ovadya's committed plea for platform democracy, through Citizen Assemblies. A useful taxonomy is the Content & Jurisdiction Program - Operational Approaches by the Internet & Jurisdiction Policy Network. In a recent paper Rachel Griffin points to the existence of two approaches to alleviate legitimacy deficits of platform decision-making, a "multistakeholderist response to increase civil society's influence in platform governance through transparency, consultation and participation" and a "rule of law response" extending "the platform/state analogy to argue that platform governance should follow the same

rule of law principles as public institutions". Most recently the Behavioral Insights Team cooperated with Meta and organized a deliberative assembly for user integration into social media decision-making at scale. These more recent developments are illuminating in that they point to a widespread discontent with platform rule (and platform rules).

– A major social network has created an Oversight Board to help with content decisions and algorithmic recommendations. A gaming label is experimenting with player councils to help programmers make exciting choices. German public television's advisory council wants to create a people's panel to ensure more input into programming decisions. The world's largest online knowledge platform has, since its inception, let users (and user-editors) decide upon content-related conflicts. All of these examples share one fundamental goal: ensuring that decisions on communication rules, for people and/or mediated through algorithms, are better, more nuanced, and considered more legitimate through broader involvement.

– What is the solution to the challenges of socially reconnecting platform decisions and practices and public values? Do we need an independent and pluralistic body that has advisory or binding decision-making powers, consisting of representatives of users, or of governmental and civil society bodies, or a combination thereof? And if yes, at which level (platform-specific, national, regional, international)? Deliberative forums have been used in different policy settings for a long time. Groups of randomly selected individuals are tasked with addressing a particular issue or question of public concern. Similar councils, usually expert-led or on the basis of multistakeholder representation, have existed for traditional media, especially public service media. However, platform councils – councils advising social media platforms – are a more recent phenomenon.

– The Oversight Board is a global, currently platform-specific, quasi-judicial and advisory platform council. The board is staffed with representatives from academia and civil society who, on the basis of Meta's community standards and international human rights standards, assess Meta's decisions on the deletion of user-generated content on its Facebook and Instagram platforms.

– The Oversight Board may be seen as the first attempt to partially open up the decision-making system of a commercial platform to the "outside"; but to form and position this "outside" in a way that not only promises systemic improvements for users, but actually triggers them beyond the individual cases remains a challenge. The Councils of other companies, including TikTok (which set up several regional "councils") and Twitter (which disbanded its Trust and Safety Council after changes in ownership), have not been institutionally strong enough to confirm a global trend. It is therefore upon scholars to develop and provide innovative concepts for a democracy-friendly design of platform councils and to innovate participatory ideas.

The role of platform councils

– The key added value of Social Media Councils is to provide more legitimacy to the rules and algorithmic practices of platforms. Any form-related decision has to follow this function.
– Our research has confirmed that no single solution to recoupling online spaces to public values exists. Indeed key questions that have to be asked and answered in the development of social media councils relate to:
  – the relationship of social media councils to the platform and modes of enforcement of its decisions,
  – the geographical scope of SMCs: local, national, regional, global,
  – the scope of SMCs: industry-wide or specific for each platform,
  – the set-up of SMCs: by a platform? through a multi-stakeholder process? by states? as a self-regulatory approach?
  – its members/involvement: who constitutes the SMC? Users? Experts? Citizens?

- its focus: individual content regulation decisions, the design of algorithms? General human rights policy?
- procedure: how are decisions arrived at?
- funding: how to guarantee independence?
- risks: how to ensure that SMCs are not coopted by bad actors?
- culturally and geographically contingent factors: Which public values are relevant in the context of hybrid platform governance? Which conditions need to be met for the SMCs to be accepted in regional/local contexts?

- Drawbacks exist. They include a weakening of state regulators, confusion of responsibility, ethics-washing, a normative fig-leaf effect, and a globalist approach to speech rules that is non-responsive to local and regional practices. There is a lack of a singular approach to improving the legitimacy of platform rules and their algorithmic governance of information and communication flows.
- Internal SMCs are implemented by platforms on their own terms to strengthen their internal processes with specific procedural safeguards. The Meta Oversight Board is an example of such an internal SMC. By contrast, external SMCs are introduced externally and can be either voluntary or obligatory. The broader a membership of Social Media Council is, the more legitimate its decisions might appear to be. However, there are substantial trade-offs to consider when incorporating too many non-experts into the assessment of rules and algorithmic practices. The need for access to specific information about a platform's operations raises issues of confidentiality.
- The new European digital law packages include references to civil society integration into the national compliance structures of the Digital Services Act without specifying how exactly this process should be implemented.
- The German national implementation law includes a reference to a cross-platform social media council to be set up by parliament to advise the national Digital Services Coordinator.

The potential impact of platform councils

- While the basic conception of platform councils – more inclusivity in decision-making and communication space-as-product design – is relatively clear, there is limited evidence, based on the studies contributed by fellows within the project, how to best construct a platform council.
- The most effective approach seems to be one of regulated self-regulation which combines commitments by states to develop a normative framework and engages platforms' interest to meet compliance obligations. Platform councils could be established at different levels in different set-ups in order to iterate attempts to increase the legitimacy of platform rules and algorithmic recommender orders. The impact of the various approaches should then be compared and contrasted and the most effective approach should be baselined for future implementation.
- A key concept to serve as guide in the development of social media councils is that of multistakeholderism. To effectively address both global and local challenges and interests, SMC governance structures must be tailored to meet the needs of all, including non-users and non-onliners.

The design of platform councils

- To ensure responsiveness to citizens' concerns resulting from participation, channels must exist for citizens' concerns to be heard, and threats to human rights must be considered. Machine-learning tools can be used to identify recurring speech conflicts. National social media councils may be more effective in addressing specific issues that are unique to a particular country, such as local laws and regulations, cultural sensitivities, and political contexts. They may also be more accessible

to marginalized groups within that country who may not have the resources to participate in regional or global councils.

- Regional social media councils may be useful in addressing issues that affect multiple countries within a region, such as language barriers, regional political and cultural contexts, and shared challenges. Regional councils may also provide a platform for regional coordination and collaboration on issues related to social media governance.

- Global social media councils may be necessary to address issues that transcend national or regional boundaries, such as cross-border hate speech, disinformation campaigns, and the influence of global social media platforms on local cultures and politics. A global council may also provide a platform for global coordination and collaboration on social media governance issues. Ultimately, the decision of whether to establish a national, regional, or global social media council should depend on the specific context and objectives of the council. A multi-level approach that involves collaboration and coordination among different levels of councils may also be necessary to effectively address the complex and multifaceted challenges of social media governance. We can learn from broadcasting councils which partially perpetuate prevailing power dynamics in society, with underrepresentation of socially marginalized groups such as the poor, disabled people, religious minorities, PoC, LGBTIQ*, and a relatively high average age of members. Furthermore, men outnumber women and people of other genders in these councils.

- Expertise matters: Incorporating technical expertise into platform council decision-making requires balancing the need for pertinent information, potentially from someone with ties to the online service being evaluated, with the importance of ensuring the independence of the expert's perspective.

- To address these gaps, a multistakeholder community is required, which brings together stakeholders from various sectors to fill the many research and resource gaps. If Social Media Councils are not properly empowered, they may end up being composed of non-diverse individuals who provide non-binding suggestions to the predominant companies on how they can superficially conform to human rights constraints. However, with adequate resources and authority, and a focus on inclusivity and promoting marginalized and minority groups, they have the potential to compel platforms and regulators to address systemic inequalities, create safer and more welcoming social media environments, and establish more egalitarian forms of social media governance.

The way forward

- This project has focused on investigating the potential of platform councils as instruments and forums for more digital democracy. Especially when social media councils are established with the goal of enabling an appropriate assessment of individual pieces of contents, the construction runs into representation problems that are almost impossible to solve, because societies have become so complex that it is hopeless to include all cultural, social and other perspectives in a single body. This is already a problem with national broadcasting councils, since there are hardly any associations left that represent large parts of society. Much attention was paid to inclusivity in the design of the Meta Oversight Board, but it is still criticized in this respect.

- Meta has introduced a novel approach for decision-making in the development of apps and technologies known as the Community Forum. The concept of community forums and online citizen assemblies could be extended to the metaverse. This creates new opportunities for participatory democracy and deliberative decision-making processes.

- The best way forward is to integrate the interests of diverse stakeholders, iterate models of social media councils and to innovate new models of recoupling public interest and private orders in hybrid communication spaces. There can be value in not waiting for the perfect solution of one

council but to try out various models in parallel, creating a regulatory system in which these observe each other and hold each other to account.

– Against this background, we want to propose an actor that can solve a specific problem for which no convincing solution exists so far. This is not a matter of assessing individual decisions on content, but rather of the level of rulemaking by platforms, and there specifically of rules that affect public interests in a specific way and are therefore in particular need of legitimation in the sense mentioned above. Building on substantial previous work, we propose to establish a commission dedicated to furthering the application of rule of law standards, of democratic values and of human rights on the Internet, in particular within platforms.

– The commission is focused on ensuring human rights online. It can be modeled after the European Commission for Democracy through Law ("Venice Commission"), an independent consultative body within the Council of Europe systems that provides expertise and conducts on issues of constitutional law and democratic institutions with a focus on good practices and minimum standards.

– Its main task is to focus on policy questions, not on individual pieces of content, and on public interest questions. These include, particularly, exceptions for public figures, inauthentic behaviour, rules for media content and algorithmic recommendations in light of content diversity. These areas are clearly not to be regulated by platforms alone, but should neither be regulated through state rules.

– The Commission can build on, and be staffed by, a diverse set of experts from selected organizations and institutions established in the platform governance field, including research centers and networks of experts.
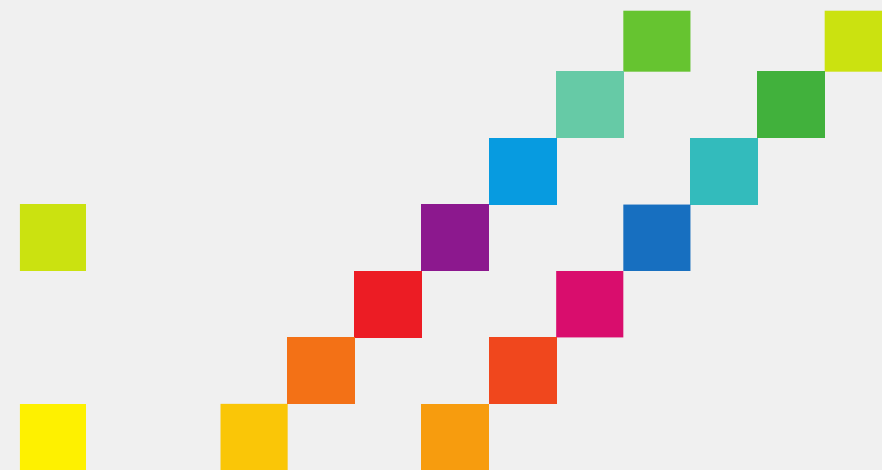
MATTHIAS C. KETTEMANN, ANRIETTE ESTERHUYSEN AND JOSEFA FRANCKE (EDS.)

Platform://Democracy

Research Report Africa

PLATFORM://DEMOCRACY

Platform://Democracy

Perspectives on Platform Power, Public Values and the Potential of Social Media Councils: Research Report Africa

edited by Matthias C. Kettemann, Anriette Esterhuysen and Josefa Francke,
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

LEIBNIZ INSTITUTE
FOR MEDIA RESEARCH
HANS-BREDOW-INSTITUT

ALEXANDER VON HUMBOLDT
INSTITUTE FOR INTERNET
AND SOCIETY

STIFTUNG
MERCATOR

Contributors

Name(s)	Affiliation
Anriette Esterhuysen	Association for Progressive Communications
Khadijah El-Usman	Paradigm Initiative
Nashilongo Gervasius	Namibia University of Science and Technology, London School of Economis
Grace Githaiga	Kenya ICT Action Network
Kuda Hove	Independent researcher
Tomiwa Ilori	Centre for Humna Rights, University of Pretoria
Ephraim Percy Kenyanito	Queen Mary University of London, University of East London
Thobekile Matimbe	Paradigm Initiative
Berhan Taye	Independent researcher
Emmanuel Vitus	AW Free Foundation

Table of Contents

Introduction to the Research Clinic Africa

Anriette Esterhuysen
ASSOCIATION FOR PROGRESSIVE COMMUNICATION, JOHANNESBURG, SOUTH AFRICA

Social media has transformed the character, scope, and scale of public discourse and political participation in Africa. In many respects the promise of digitalisation for social and economic development in the region has not been realised, and the average overall internet penetration is still well below 50%. Differences in internet penetration between regions and countries are dramatic, for example, in January 2022 it ranged from 7.1% in the Central African Republic to 84% in Morocco.[1] Insufficient availability of infrastructure and the high cost of data - Most African internet users connect via mobile phones - and devices continue to limit people's ability to access the internet. But in spite of these barriers, Africans have taken to social media with great enthusiasm. Patterns of use vary from other parts of the world in some respects, for example, statistics for 2022 indicates that average WhatsApp use in Africa is 35% higher than the global average.[2] Facebook is the most widely used platform, followed by YouTube, Twitter and Instagram.[3]

More Africans use social media to communicate than any other form of communication. Phone calls are expensive, and so is Short Message Service (SMS). Short calls or messages sent over social media platforms cost less, even though the cost of data in Africa is still comparatively high. A recently published commercial media monitoring survey found that, from among more than 170 participating countries, "African countries rely more on social media as a marketing tool (74%) compared to the overall results of other regions (52%)."[4] An Afrobarometer survey released in early 2022 indicates that use of digital sources for news had nearly doubled since 2015 with "more than four in ten adults across 34 surveyed countries reporting that they turn to the Internet or social media at least a few times a week for news".[5] Reliance on social media for news is reinforced by widespread lack of confidence in print and broadcast media.[6] Twitter is used widely for political conversations. African leaders tend to be ambivalent, on the one hand using social media themselves, and on the other hand feeling threatened and readily resorting to shutdowns during elections or protests.

Social media has become part of the fabric of daily life in Africa, yet African internet users have had very little influence on how it is governed. At national level many governments have introduced poorly thought-through regulations under the auspices of targeting harmful content and mis- and disinformation. However, these laws have generally been more effective at stifling online freedom of expression than in achieving the stated objective of curbing false information.[7] Nor have Africans had

---

[1] https://www.statista.com/statistics/1124283/internet-penetration-in-africa-by-country/

[2] State of Social Media for Africa 2023, Meltwater, January2023. https://meltwater.cdn.prismic.io/meltwater/91f084cd-7cb6-45a2-b072-4e319783e585_StateofSocial_Africa.pdf

[3] https://www.itnewsafrica.com/2022/07/top-5-most-popular-social-media-platforms-in-africa/

[4] https://www.bizcommunity.com/Article/196/669/235086.html

[5] https://www.afrobarometer.org/publication/ad509-promise-and-peril-in-changing-media-landscape-africans-are-concerned-about-social-media-but-opposed-to-restricting-access/

[6] In their 2022 World Press Freedom Index, Reporters Without Borders rated press freedom in only seven African countries as being satisfactory. https://rsf.org/en/index

[7] Misinformation Policy in Sub-Saharan Africa: From Laws and Regulations to Media Literacy, Universeity of Westminster Press, 2021. https://www.uwestminsterpress.co.uk/site/books/m/10.16997/book53/

much influence on how global social media platforms are governed, as pointed out in most of the contributions below.

But this lack of influence is not one dimensional. The perspectives on platform power, public values, and the potential of social media councils and other forms of platform governance discussed in the eight short papers below provide examples of how Africans can have more agency in the governance of social media platforms.

In *Rethinking Platform democracy in Africa: A Nigerian case study,* Khadijah El-Usman argues that social media platforms have acquired legitimacy in Nigeria but as they are based outside Africa and African users have no say in how they are governed, the nuances of indigenous languages, regional and political contexts, and larger interests of African users are often missed. On the other hand, where governments seek to enforce their own platform governance rules they often conflict with national and international human rights principles. Therefore, there is a need to introduce meaningful participation by African users in Social Media Councils, and combine global principles with national and regional regulation that is attuned to local contexts. She explores how existing regional regulation in West Africa can be used to resolve platform related complaints.

Nashilongo Gervasius Nakale examines decisions of the Meta Oversight Board's decisions in the African context and concludes that it applies universal laws to African realities, without making a serious effort to apply applicable national or regional regulatory frameworks. Consultation with local stakeholders, when there is any, is limited. Global rights frameworks and values are privileged at the expense of local nuances and contexts. An alternative approach would be to pay more attention to national legislative and human rights frameworks, and regional African human rights instruments, and convene more regional consultations. This would complement the existing oversight process. By involving local and regional bodies including civil society organisations, regulatory bodies, and human rights oversight mechanisms at the African Union level the Oversight Board's decisions would become more respectful of national democratic laws and processes.

Kuda Hove points out the same problem in his paper *A Proposal on Voluntary Social Media Councils.* To overcome that, he proposes that local regulatory models used for the print media can be modified and applied to social media platforms. He proposes the idea of voluntary national level social media councils which could consist of civil society actors, academics, users and representatives of special interest groups and minorities, to close the gap between the global and the local, and provide a legitimately independent, participatory, and transparent way to ensure user participation in content moderation and the regulation of social media platforms in each country.

Emmanuel Vitus' *Platform Democracy, one size does not fit all: the case of GhanaWeb* also focuses on the tension between the global standards used by social media platforms and the need to defer to local contexts when moderating content. He presents GhanaWeb an example of balancing freedom of speech for users and global trends to moderate online content perceived as potentially harmful. It engages in several phases of technical and human review in order to identify, assess, and take action against content that potentially violates its Community Standards through using local languages. The case of GhanaWeb shows how local platforms in Africa are struggling, and in this case succeeding, to build a transparent and strong moderation system to enhance user expression/experience and also keep the platform safe and democratic. There is a need to balance moderation with censorship, publish content guidelines and policies, disclose moderation practices and appeals process, and disclose algorithms.

In her paper on *Elections and social media platforms in Kenya* Grace Githaiga argues that social media platforms failed to uphold their commitment to maintain election integrity and prevent the proliferation

of harmful content during Kenya's 2022 elections. Civil society groups felt that moderators did not understand regional languages and did not take into consideration cultural and societal settings. There is a need for mechanisms to encourage fact-checking, combat hate speech, and engage stakeholders, including citizens and users, to report offenses, and to participate actively. Civil society must be strategic in engaging media platforms and point out local challenges. There is a need for international partners to make resources available resources to Global South Civil Society organisations in an effort to support their work in campaigns of engagement.

Ephraim Percy Kenyanito's article, *Unpacking the Tensions: A Comparative Analysis of DNS Abuse Mitigation and its Impact on Business and Human Rights* looks at DNS abuse in different countries, which poses a significant threat to human rights, particularly the right to freedom of expression. While DNS abuse mitigation measures are necessary, they must be consistent with international human rights law and not result in censorship or violate other human rights. Human rights considerations need to be integrated into DNS abuse mitigation measures, together with promotion of transparency and accountability in the implementation of these measures, and regular assessments of registry policies. By taking these steps, DNS abuse cane be mitigated while protecting human rights.

In *Centering Victims is Imperative for Effective Remediation in Platform Governance* Thobekile Matimbe calls on social media platforms to adopt a victim-centred approach as key in interventions that promote human rights. They should be transparent about how they address human rights and the remedial steps they take to address victims of harmful online content. Consulting with victim support groups and clinical psychologists is part of engaging different stakeholders in addressing victim needs. Platforms need to cooperate with other accountability mechanisms within the different contexts and jurisdictions where harmful content originates.

Tomiwa Ilori tackles what is in effect the overall challenge of platform governance in Africa in his article *Building and strengthening rights-based social media platform governance in Africa through national human rights institutions.* He points out that the laws, policies and processes made by African governments to regulate online harms are often at variance with international human rights standards while social media platforms are distant from the contextual realities required to regulate online harms in African countries. National Human rights institutions (NHRIs) can provide a link here as they have a key responsibility to ensure that key international human rights instruments are considered and implemented in local contexts. However, African NHRIs face a number of challenges that could make playing these roles difficult. They need to become involved in ensuring a rights-based approach to social media platform governance, build and strengthen the capacity of duty-bearers to meet their obligations while also encouraging rights holders to claim their rights. Some of the ways through which such a role can be played include strategic collaboration, policy advocacy, capacity building and radical participation.

Four themes cut through all the papers. First, the tension between using global standards as a basis for decision-making on whether content should stay or go, versus taking local contexts, cultures, values, and regulatory environments into account. Second, and linked to the first tension, is the difficulty of finding a balance between recognition and respect of local authorities and values without endorsing censorship, bias, or undue restrictions on human rights. Third is the challenge to achieve more and deeper, engagement from the people and institutions in Africa on how global governance frameworks are designed and applied. This involves creating more awareness of social media councils in the region, and investing in increasing African participation in their processes. The fourth theme is the interplay between international, regional, and national governance structures and standards, and it raises the question of how, over the longer term, platform governance can either strengthen or undermine

transparent, accountable, and human rights-based content governance in the region. A key example would be the African Commission on Human and People's Rights Declaration of Principles on Freedom of Expression of 2019, a document intended to guide the behaviour of states region at country level. It was developed through extensive consultation with African media, internet and human rights organisations across the continent. It articulates international human rights standards that apply to the internet in from an African perspective for the African context. By bypassing this regional instrument, social media councils undermine African efforts to promote respect for human rights online.

What emerges from the papers is that existing national and regional institutions and governance arrangements (including self-governance and co-governance) can serve as a source for greater platform democracy, and that making use of these can serve to strengthen them. Ignoring them can result in consolidating the counter-productive notions that democracy and human rights are "un-African".

Rethinking Platform democracy in Africa: A Nigerian case study

Khadijah El-Usman
PARADIGM INITIATIVE, LAGOS, NIGERIA

This paper explores how one determines legitimacy and differentiates it from democratic platform governance, which standards platforms can adopt in terms of governance to make themselves democratic and legitimate, especially in decision making, rule enforcing and decision reviewing. It further explores whether rule enforcing and decision reviewing be community-based, national-based or regional, considering the possibility of authoritarian governments trampling on rights. The paper analyses these questions from a Nigerian perspective.

Introduction

In 2020, Nigerian youth took to the street to protest police brutality after a series of reports emerged of extrajudicial killings orchestrated by a special unit of the police force.[1] This protest was organised around the country using Twitter. Via this platform, the youth mobilised, canvassed for and distributed supplies and donations while also spreading awareness in a manner that's been unprecedented since the nation's return to democracy.[2] It was a pivotal moment as platforms, especially social media, have become synonymous with providing individuals with unprecedented self-determination through freedom of expression, participation in the democratic process and its public square nature of discussion and information dissemination. During this time, credible news sources quoted social media posts and live streams, many of said news channels were eventually fined for doing so, basing entire news stories on decentralised reports from users of these digital channels.[3]

More upheavals followed this digital revolution. In early 2021, the president of Nigeria, via an official account, put out a tweet that was mass reported for being abusive and tribalist. Twitter ultimately took the tweet down for failing to adhere to its content moderation rules. Consequently, there was an uproar originating from members of government and the ruling administration at the time. Despite a considerable number of cheers from Nigerian users of the microblogging site, the actions of Twitter's leadership were seen by some as an affront to Nigeria's sovereignty. Many have since linked these events to the infamously named "Twitter ban" in Nigeria, which ran from June 2021 to January 2022.[4]

These events have since become catalysts for Nigerians to look deeper into the questions of platform legitimacy and democracy. They've also necessitated questions about the standards to be adopted by platforms in decision-making and what models will work best. This paper discusses considerations

---

[1] Chiamaka Ozulumba, What Led to #EndSARS Protests?, Thisdaylive, https://www.thisdaylive.com/index.php/2021/10/20/what-led-to-endsars-protests/#:~:text=EndSARS%20protests%20began%20as%20a,citizens%20and%20violated%20their%20rights.Accessed 25th February, 2023

[2] Emmanuel Elebeke, #ENDSARS Coverage: Why we fined AIT, Channels, Arise TV, N3m each — NBC. Vanguard

[3] https://www.vanguardngr.com/2020/10/endsars-coverage-why-we-fined-ait-channels-arise-tv-n3m-each-%E2%80%95-nbc/#:~:text=They%20include%3A%20Channels%20Television%2C%20African,General%20of%20the%20Commission%2C%20Prof Accessed 25th February, 2023

[4] Aljazera, Nigeria ends its Twitter ban after seven months, https://www.aljazeera.com/economy/2022/1/12/nigeria-ends-its-twitter-ban-after-seven-months#:~:text=Nigeria%20ends%20its%20Twitter%20ban,Social%20Media%20News%20%7C%20Al%20Jazeera, Accessed 25th march,2023

revolving around platform legitimacy among Nigerian citizens and how best platform democracy can be achieved; it further explores stakeholders to be involved in platform decision-making and the geographical scope of these decisions.

Defining legitimacy/democracy

Platform legitimacy and democracy, although related, are separate concepts.

Legitimacy refers to the recognition and acceptance of authority or power by the people subject to that authority.[5] It is based on the idea that the exercise of power should operate through consent and be seen as fair and just by those subject to it.

Democracy, on the other hand, is a form of government in which power is held by the people, either directly or through elected representatives. It is based on the idea that the people should have a say in how they are governed and that the government should be accountable to the people.[6]

Relating both concepts to the context of platforms, legitimacy would translate to a platform being wholly accepted by users from its content guidelines, terms and conditions to simply its existence as a part of daily living. Similarly, platform democracy would constitute the extent to which people using this platform are involved in exercising that power and have a say in its governance.

In Nigeria, a certain level of legitimacy has been attained by platforms. The most significant indicator of this has been citizen pushback against the government when it sought control over these platforms or tried to dictate what kind of content should be published and disseminated. In the wake of a Twitter ban, citizens fought and protested at various levels to have the platform restored. Multiple legal cases were instituted across several courts, and many sought to circumvent the block using Virtual Private Networks (VPNs) to access the platform. Most citizens sided with the platform's decision to take down the President's tweet. The action was notably viewed as another avenue for the electorate to exercise its power against the most powerful man in the nation. This show of strength indeed showed the platform had gained some form of legitimacy.

Democracy on platforms

Although platforms have gained acceptance, there is little evidence to suggest they are committed to being democratic. While social media councils - the bodies in charge of platform advice and adjudication with content moderation policies - exist, there is still room for a lot more improvement. While Meta's Oversight Board includes greater diversity than other major social media platforms, it's still not quite democratic. Meta chose the initial board members, and when their terms are up, they will choose their replacements.[7]

Wider questions regarding the ultimate motivations of platform owners and operators remain. With the biggest platforms being largely based outside Africa, their African users have not had a say in how they are governed. Consequently, the nuances of indigenous languages, regional and political contexts, and

---

[5] Ian Hurd. Legitimacy,
https://pesd.princeton.edu/node/516#:~:text=Legitimacy%20is%20commonly%20defined%20in,toward%20the%20rule%20or%20ruler, Accessed 15th march,2023

[6] Stanford Encyclopedia of Philosophy, Democracy, https://plato.stanford.edu/entries/democracy/#DemoDefi, Accessed 15th march,2023

[7] Oversight board bye-laws, https://transparency.fb.com/sr/oversight-board-bylaws-2023, Accessed 11th March, 2023

larger interests of African users (beyond revenue for the platforms and ever-increasing user bases) are often missed.

The emergence of these platforms, in disrupting the distribution of power and information in varying degrees, poses a novel set of challenges to state authority. It affects state sovereignty and in response, the Nigerian government has sought to make many attempts to control this space. The current administration has attempted to institute various forms of restrictive legislation, including a hate speech bill, a social media bill and a code of practice for computer service platforms. Each has been put forward in a bid to wield powers that would ensure a certain style of ownership of the decision making processes of the platforms. Some areas of concern have been mis/disinformation, hate speech, unlawful content, and content taking down time among others. Although these seem on the face of it to be beneficial interventions for citizens, the bills often have provisions that can be subject to abuse by government. One of such provisions can be found in the Hate Speech Prohibition Bill 2019[8] which states that anyone found guilty of hate speech is liable to 3-5 year imprisonment while the social media bill[9] was to allow the government to direct the internet to be shut down, restricting the rights to freedom of expression and access to information. The provisions do not come off as commensurate to the crime, and many feared they would be used to target any government dissidents, as has been witnessed with the Cybercrimes Act 2015[10] where journalists were arrested for reporting unfavourable against the government of the day. Although certain provisions in these bills could be seen as positive, any user of the Internet could fall on the wrong side of the proposed bill, as "falsehood" could have relative meanings.

These attempts were summarily condemned and protested[11] against by distrusting citizens especially because the synergy sought for between the platform and the state seemed to propose attention being paid to only government interests.

After the Twitter ban, the government presented Twitter with a raft of terms including being registered in Nigeria, opening a Nigerian office among others, citing them as conditions that would favour the people and facilitate its return. Of the conditions given by the Government, providing a direct channel to the Partner Support Portal for law enforcement, government officials and Twitter staff to manage prohibited content that violates Twitter community rules was met with the most distrust due to established patterns.

Where the government seeks to be involved or enforce its own platform governance rules, it is often conflicting with national and international human rights principles. Such demands far too often aim to fit a political agenda. On the other hand, when platforms enforce their decision making process or act slowly on certain reports, it often lacks indigenous contexts such as missing hate speech or abusive language said in indigenous language. Therefore, both outcomes leave considerable possibilities for harm to be perpetuated.

Stakeholder involvement in decision making

---

[8] Hate speech (prohibition) bill 2019, HB 246, https://placbillstrack.org/upload/HB246.pdf, Accessed 11th March, 2023

[9] Protection from Internet Falsehood and Manipulation Bill 2019, https://guardian.ng/wp-content/uploads/2019/11/Protection-from-Internet-Falsehood-and-Manipulation-Bill-2019.pdf, Accessed 9th March, 2023

[10] Premium Times. Journalist arrested, charged under cybercrime law in Nigeria, https://www.premiumtimesng.com/news/top-news/318321-journalist-arrested-charged-under-cybercrime-law-in-nigeria.html?tztc=1, Accessed 13th March, 2023

[11] John Akubo, Tina Abeku (Abuja), Opeyemi Babalola (Lagos) and Akin Alofetekun (Minna), Protests against hate speech, social media bills rock Abuja, Lagos, https://guardian.ng/news/protests-against-hate-speech-social-media-bills-rock-abuja-lagos/, Accessed 10th March, 2023

Social media councils have gained popularity with platforms.[12] They are instituted with the aim of serving advisory and judicial roles in decision making. However, a few flaws are rife in their practice with regard to Africa. First, there is the fundamental flaw of attempting to use one brush to paint all canvases, which is to say that using one global council often based in Europe or America, is unsuitable. Second, is the self regulatory, voluntary implementation model of most social media councils[13] and that existing Social Media Councils lack meaningful participation by African users/citizens.

Establishing a regional or national Social Media Council with stakeholder participation could be a solution and entrenching such a practice will assist platforms to attain increased legitimacy among the people or users. There are however various demographics of stakeholders to involve. Government agencies in Nigeria have adopted a model of stakeholder participation in decision making, that could be used as role model for platform governance. Various demographics and bodies are often represented on the agency's governing board. [14] One of such agencies of note is the National Information Technology Development Agency (NITDA), whose board is statutorily mandated to consist of government agencies in ICT, representatives of computer service bodies, academic staff, representatives from six geopolitical zones and others. It is, however, often criticised for not reaching all the necessary stakeholders.[15] A Nigerian-national based social media council should have the compulsory addition of the human rights commission, civil society organisations and youth participation. This consideration will take advantage of the massive youth participation on platforms, ensuring that a critical mass of users have their concerns and interests represented via relevant and inclusive governance.

National and regional platform governance: the ECOWAS and OHADA models

The question of whether a national or regional approach should be adopted in platform regulation in Africa is a complex one. It is easily made more complicated when viewed in the light of the operation of social media councils. Arguments can be laid out for both approaches.

A national approach to platform governance could allow for more targeted and situation-specific rules that take into account each country's unique needs and challenges. It could also encourage people to take more ownership of and part in making policies and practices, which could help build trust and legitimacy in the regulatory process.

On the other hand, a regional approach to platform regulation could promote greater harmonisation and consistency in the regulatory landscape across different countries in Africa. This could help reduce confusion and fragmentation, and also provide a more level playing field for platform companies and users across the region. A regional approach could also facilitate greater collaboration and information sharing among clusters of countries in addressing common challenges and issues related to social media.

---

[12] Article 19, Social Media Councils: One piece in the puzzle of content moderation, https://www.article19.org/resources/social-media-councils-moderation/, Accessed 10th March, 2023

[13] Riku Neuvonen, Internet and Self-Regulation: Media Councils as Models for Social Media Councils?, https://graphite.page/gdhrnet-platform-response/assets/documents/GDHRNet-BestPractice-PartV-2.pdf, Accessed 5th March, 2023

[14] National Information Technology Development Agency Act 2007 Act No. 28 Published In The Federal Republic Of Nigeria Official Gazette No. 99 Vol. 94 Lagos 5th October 2007, https://nitda.gov.ng/wp-content/uploads/2020/11/NITDA-ACT-2007-2019-Edition1.pdf, Accessed 3rd March 2023

[15] Sultan Quadri, Nigeria's House of Assembly adjourns public hearing on controversial NITDA bill, https://techcabal.com/2022/12/24/nigerias-house-of-assembly-adjourns-public-hearing-on-controversial-nitda-bill/, Accessed February 29th 2023

An argument significantly in favour of the regional approach is the propensity of African governments to become authoritarian. For instance, in 2021, 12 countries barred internet access for their citizens at least 19 times on the continent.[16] Thus, where a government becomes authoritarian or espouses authoritarian principles, platforms must return to the role of respecting human rights with the regional body being able to override the national body.

Yet, Africa's complexities, as exhibited by its regional classifications from geography to language and norms, could pose a challenge. This diverse range of interests could make it harder for the entire continent to agree on one form of platform regulation. Meeting halfway, Africa's regional economic communities could serve as orientation, and in Nigeria's case, this would be the Economic Community of West African States (ECOWAS)[17].

The ECOWAS comprises three arms of governance, namely, the Executive, the Legislature and the Judiciary. All three arms in concert, and within them is an established structure that sees to the application of Community laws, protocols and conventions. ECOWAS has unfortunately struggled with adopting and implementing its own laws within the region because although ECOWAS laws are adoptable at a regional level, most member states will need it to be integrated into national laws by their own parliaments.[18] This provision is also in the ECOWAS Revised treaty[19] which states that

"Each Member State shall, in accordance with its constitutional procedures, take all necessary measures to ensure the enactment and dissemination of such legislative and statutory texts as may be necessary for the implementation of the provisions of this Treaty."

This is a difficult process to complete as ECOWAS states often make an argument for state sovereignty. This extends beyond ECOWAS laws to every other treaty. An example of this is the ruling of a Nigerian court in the case Abacha v Fawehinmi, which examined the status of the African Charter on Human and Peoples´ Rights that was domesticated in Nigeria by legislation. The Court held that the Nigerian Constitution is superior to the Charter. However, that is not to say that none of the treaties have managed to be adopted and integrated into national law, essentially practising a dualist[20] system of international law.

A regional body that has managed to achieve this within Africa is the Organization for the Harmonization of Business Law in Africa (OHADA)[21], an intergovernmental organisation for legal integration. It was established by the Treaty of 17 October 1993 signed in Port Louis (Mauritius), as revised on 17 October 2008 in Quebec (Canada) by francophone African countries which Nigeria is not a part of.

---

[16] Access now, Internet shutdowns in 2021 report: resistance in the face of blackouts in Africa, https://www.accessnow.org/internet-shutdowns-africa-keepiton-2021/, Accessed February 20th, 2023

[17] ECOWAS, member states, https://ecowas.int/member-states/, Accessed 27th March, 2023

[18] sec 1(2) of the Constitution of Ghana, see also sec 12(1) of the 1999 Constitution of the Federal Republic of Nigeria.

[19] Article 5(2), Economic Community of West African States (ECOWAS), Revised Treaty of the Economic Community of West African States (ECOWAS), 24 July 1993, available at: https://www.refworld.org/docid/492182d92.html [accessed 30 March 2023]

[20] Carolyn A. Dubay, General Principles of International Law: Monism and Dualism,International Judicial Academy, Washington, D.C., http://www.judicialmonitor.org/archive_winter2014/generalprinciples.html#:~:text=Under%20a%20dualist%20model%2C%20there,and%20its%20citizens%20or%20subjects., Accessed 29th March, 2023

[21] Organisation for the Harmonization of Business Law in Africa,General Overview, https://www.ohada.org/en/general-overview/, Accessed 29th March 2023.

What makes the model unique is the practice of the monist[22] system of international law seen in the provisions of Article 10 within it, the Uniforms Acts, which constitute the Ohada Laws, are directly applicable and binding on member states. The Uniform Acts prevail in the event of conflicting provisions of any domestic laws of member-states.

Article 10 states as follows:

"Uniform Acts are directly applicable and overriding in the contracting states notwithstanding any conflict they may give rise to in respect of previous or subsequent enactment of municipal laws."

The Ohada Treaty and laws offer a unique model for the harmonisation of business laws as a starting point for the harmonisation of other laws within ECOWAS states. The Ohada Initiative has created a body of harmonised business laws in its member-states, which have, in consequence facilitated business transactions to a large extent. A similar adoption with ECOWAS or its member states joining OHADA in a bid to implement social media regulation will facilitate platform legitimacy, democracy, and trust within the region. This will not be the first time it has been suggested that ECOWAS adopts the OHADA model or simply joins it. Various authors[23] have posited the same citing the merits of such a merger.

In considering both approaches to platform regulation, what will work best will be a tiered approach. This is because, in practice, both national and regional approaches are likely to be needed in order to effectively regulate platforms in Africa. Countries in the region face unique challenges and circumstances that require tailored and responsive regulatory frameworks while also requiring regional cooperation and coordination in order to address transnational issues such as hate speech, disinformation, and online harassment.

Within a tiered approach, the national body (social media council) is the immediate body for decision making while the regional body will exist as a decision reviewing body. This can be accomplished through the adoption of a uniform legal framework by the new established platform regional body that is based on principles of transparency, accountability, and respect for human rights and must be developed in consultation with a wide range of stakeholders from across the region. States have been known to come together when their interests align, such as in the case of the African Continental Free Trade Area[24], and platform regulation might be one of those places. One key word that is commonly used in both Ohada and Ecowas Treaties is 'integration'. Integration has many facets, including, economic, political, social, geographical, and legal. Any kind of integration must be predicated on a legal framework.

Recommendations

User participation in platform decision making and norm setting becomes a priority for social media platforms, taking into consideration national and regional contexts long before establishing social media councils.

---

[22] n17.

[23] http://www.nigerianlawguru.com/articles/international%20law/OHADA%20TREATY%20AND%20LAW%20BY%20ANGLOPHONE%20STATES.pdf

[24] Tralac,Status of AfCFTA Ratification, https://www.tralac.org/resources/infographic/13795-status-of-afcfta-ratification.html, Accessed 30th March, 2023

In establishing social media councils the adoption of a two tiered system of national and regional platform regulation where the national system covers national decision making and regional body covers decision reviewing should be considered. A tiered approach should be adopted using a regional legal framework for platform decision making and accountability, where a national body exists on decision making, while the regional body takes on decision reviewing using the monist system of international law, where decisions of the regional body surpass that of the national one.

The composition of the national and regional social media councils should comprise various stakeholders and have to include national human rights organisations, civil society organisations and youth representatives.

The overarching aim of this paper was to identify the contextual factors in Nigeria's platform democracy and suggest solutions. Although solutions have been recommended, openng these assertions to further research and testing the models for social media councils practically using case studies, feasibility studies etc. is needed.

The Oversight Board's decisions in the African context

Nashilongo Gervasius Nakale
NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY, WINDHUK, NAMIBIA

The question of self-regulation of platforms through social media councils (many of whom are funded by social media companies) has made the rounds since Facebook (now Meta) announced the establishment of what is called the Oversight Board in 2018. This scenario of a global body making decisions on content takedowns at the national level raises the further question of how self-regulation operates in the context of hyper-globalization, and how this relates to democratic governance and regulation, which has different shapes in different states. Hyperglobalization can be described as "the dramatic change in the size, scope, and velocity of globalization that began in the late 1990s and that continues into the beginning of the 21st century. It covers all three main dimensions of economic globalization, cultural globalization, and political globalization."[1] Its impacts can make it difficult for a government or a country to develop democratic governance on its own terms, maintaining its national sovereignty while also achieving deeper international integration in a hyper-connected world. This is particularly challenging in sub-Saharan countries where digitalization can be linked to new forms of (de)colonisation[2] with nations struggling to maintain self-determination, local traditions, cultures, and values. In an increasingly globalized world universal values are prioritized over local values, particularly online.

Social media councils like the Meta Oversight Board serve as the final layer in the content moderation process, reviewing contested decisions made by the platform and supposedly offering a fair hearing to those users whose content or concerns are affected.[3] The legitimacy of social media councils such as the Oversight Board (OB) is established through the expertise and diversity of its members and the independence of the review process and the resulting decisions. The journey taken by the OB to arrive at either overturning or maintaining decisions made at the platform level involves a process that is supposed to consider local contexts and international human rights standards.

Looking at OB decisions on cases from sub-Saharan Africa during 2021 and 2022 this study found that in practice, their processes pay scant attention to local contexts, values, and regulatory systems. In fact, their decisions appear to almost exclusively reflect globalized values and approaches to online discourse and democracy. In the OB's review of cases, this "globalized" frame of reference takes precedence over efforts to understand the dynamics of the social environments in which the cases take place. What is even more striking is that they do not make reference to regional African human rights frameworks.

While the OB endeavors to be independent in its decision-making process, it's important to understand the challenge it faces from a theoretical perspective. One way of thinking of this challenge is to draw on the "Internet-Governance Impossibility Theorem" which proposes a political Trilemma for the World

---

[1] https://en.wikipedia.org/wiki/Hyper-globalization

[2] Danielle Coleman, Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws, 24 MICH. J. RACE & L. 417 (2019). Available at: https://repository.law.umich.edu/mjrl/vol24/iss2/6

[3] Oversight Board Bylaws. (2022). Oversight Board. https://www.oversightboard.com/sr/governance/bylaws

Economy: one cannot have hyper-globalization, democracy, and national self-determination all at once (Rodrik, 2010 p.200 as cited by Haggart, 2020 p.326).[4] Framed within the role of social media councils and their decisions, this approach states that global interconnectedness leads to limitations of democratic and self-determined solutions. A figurative example would be that a country cannot attain decisions reflecting its sovereignty based on local values including cultural and political nuances, given the digital and political interconnectedness of the world. Within the context of reviewing contested decisions, this would mean that decisions by global platforms on local matters might not necessarily be democratic nor reflective of the sovereignty of a particular state.

The Sub-Saharan African region is a region where democracy presents itself in different ways and degrees, given its diverse governance and political systems, some of which still reflect the region's colonial past.[5] Given that social power structures often mirror the state's, there is a strong likelihood that "universal" or global values might conflict with local ones. An example would be in the case of the 2022 scenario in Uganda where decisions on content based on local anti-homosexuality laws would be irreconcilable with decisions on legitimate content through content moderation based on universal values. If the OB gives precedence to these universal values, this would inevitably result in local laws being disregarded. The research done for this paper suggests this happens more often than not. The result is that the OB is not taking into account the 2019 letter[6] of the Special Rapporteur on Freedom of Expression to Facebook Founder Mark Zuckerberg that highlighted that while international human rights law would provide the board "with a set of tools and a common vocabulary for addressing and resolving hard questions around the moderation of online content" they cannot be used to solve cases as international human rights laws are originally designed to govern the relationship of state authorities with individuals and groups. In this case, the OB is neither a state authority nor does it act on behalf of the state.

This paper looks at decisions on African cases by the Meta OB over the last two years. It examines the process followed, the participation of local people and entities, and the explanations given for the decisions. The paper studies these decisions from a self-regulation and knowledge structure perspective using the Mansell & Steinmueller (2020)[7] framing that self-regulatory practices of digital platforms have the potential to make private decisions on behalf of the public. This framework is worth considering given that the Meta Oversight Board reviews decisions by the platforms that established it, even if the decisions are presented as being 'public'. Hence this paper explores the board's intricate decision-making practices and its potential to make decisions that are private in their nature on behalf of the public, a challenge highlighted by Haggart (2020) who is inclined that there are "specific implications of requiring private corporations to adhere to human rights provisions."

Oversight Board decisions on cases from Africa during 2021 and 2022

---

[4] Haggart B (2020) "Global platform governance and the internet-governance impossibility theorem" Journal of Digital Media & Policy, Volume 11, Issue Regulating Digital Platform Power, Nov 2020, p. 321 - 339

[5] Njoh, A. J. (2000). The Impact of Colonial Heritage on Development in Sub-Saharan Africa. Social Indicators Research, 52(2), 161–178. http://www.jstor.org/stable/27522501

[6] D.Kaye, (2019) Letter to Facebook CEO, "Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression" available at https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/Legislation/OL_OTH_01_05_19.pdf

[7] Mansell, Robin & Steinmueller, W.. (2020). ADVANCED INTRODUCTION TO PLATFORM ECONOMICS.

Decisions brought to the OB are published on the board's website. These are accessible in full text on the body's website depicting the type, location, the date it was published, as well as the country where the case was raised. Of the 35 decisions published by February 2023, only seven of them originated in the African region. Wong and Floridi (2022) highlight this lack of geographic diversity in case reporting from the global South when they point out that "significant geographic regions such as Sub-Saharan Africa and Central and South Asia represent[ing] just two percent of appeals".[8] Africa cases account for only about 1% of the total number of cases. This reflects little knowledge of or access to the appeal processes by users in the region.

The table below briefly highlights the decisions to remove content and their nature.

Description	Nature	Country	Region	Published	Decision
Video After Nigeria church attack	Violent and graphic content	Nigeria	SSA	Dec, 2022	Overturned
Tigray Communication Affairs Bureau	Violence and incitement	Ethiopia	SSA	Oct 4, 2022	Upheld
Reclaiming Arabic Words	Hate speech	Morocco, Egypt, Lebanon	MENA	June 13, 2022	Overturned
Sudan Graphic Video	Violent and graphic content	Sudan	SSA	June 13, 2022	Upheld
Alleged Crimes in Raya Kobo	Hate speech	Ethiopia	SSA	Dec, 2021	Upheld
South Africa Slurs	Hate Speech	South Africa	SSA	Sept 28 2021	Upheld
Shared Al Jazeera Post	Dangerous individuals & organizations	Israel, Egypt	MENA	Sept 14, 2021	Overturned

Three of the above seven decisions on African cases were overturned by the OB, while four of the decisions were upheld. Six of the seven cases stem from countries where conflict exists and/or human rights are constantly under threat. Out of the seven decisions listed in the table, five) are from Sub-Saharan African (SSA) countries, and it is these that the rest of the report concentrates on.

The investigation process

The OB upholds or overturns Meta's decisions. In doing so, the Board considers Meta's content policies and international human rights standards, particularly the International Covenant on Civil and Political Rights. This international/universal approach holds potential challenges according to Barata (2022) who opines that "international human rights law is originally designed to govern the relationship of State authorities with individuals and groups". The application of international laws in local contexts could pose further challenges according to Evelyn Douek (2021) who advocates that "International human rights laws… in content moderation should rather serve the interests of users and society rather than being co-opted by platforms to their own ends"[9]. This scenario proves the applicability of the Internet Governance Impossibility Theorem, which calls for consensus building in solving local issues at the local level, in order to achieve self -determination in democracy.

As part of the Oversight Board's procedures, an explanation and disclaimer are published alongside the cases on how decisions are prepared, the approval process, and what the decisions represent. A review of these explanations and disclaimers done for this paper indicates that, for all the African cases, the OB

---

[8] Wong, D., Floridi, L. Meta's Oversight Board: A Review and Critical Assessment. Minds & Machines (2022). https://doi.org/10.1007/s11023-022-09613-x

[9] Douek, Evelyn, The Limits of International Law in Content Moderation (October 12, 2020). 6 UCI J. INT'L TRAN'L & COMP. L. 37 (2021), Available at SSRN: https://ssrn.com/abstract=3709566 or http://dx.doi.org/10.2139/ssrn.3709566

commissioned independent research by an unnamed institute, headquartered at the University of Gothenburg in Sweden. This institutes's role is to review and substantiate the OB's findings by looking at the local context. A disclaimer on each case reviewed further states that this institute draws on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The OB is further "assisted by an advisory firm focusing on the intersection of geopolitics, trust, and safety, and technology called Duco Advisors, and Memetica, a digital investigation group providing risk advisory and threat intelligence services to mitigate online harms".

Details of the investigation processes for the five cases of focus are highlighted below. The analysis starts with the only case where a decision was overturned. The next table analyses cases where decisions were upheld. The tables include the decisions, the structure of investigations, public comments/third-party submissions as well as policy advisory statements.

Overturned cases

Case	Video after Nigeria church attack
Decision	The Board has overturned Meta's decision to remove a video from Instagram showing the aftermath of a terrorist attack in Nigeria.
Structure of Investigation	1. The user explained that they shared the video to raise awareness of the attack and to let the world know what was happening in Nigeria. 2. The Board asked Meta 29 questions, 28 of which were answered fully. Meta was unable to answer a question on the percentage of user reports that are closed without review in the Sub-Saharan Africa market. 3. The Board looked at the question of whether this content should be restored through three lenses: Meta's content policies, the company's values, and its human rights responsibilities. 4. The Board focused on three of Meta's content policies: Violent and Graphic Content; Bullying and Harassment, and Dangerous Individuals and Organizations. 5. The majority of the Board found that no policy was violated.
Public Comments/Third Party Submissions	Nine public comments related to this case are considered. One of the comments was submitted from Asia Pacific and Oceania, one from Central and South Asia, one from the Middle East and North Africa, one from Sub-Saharan Africa, and five from the United States and Canada. The submissions covered themes including the need to clarify the Violent and Graphic Content policy, and Nigeria-specific issues that the Board should be aware of while deciding this case.
Policy Advisory Statement	Content policy: Meta should review the public facing language in the Violent and Graphic Content policy to ensure that it is better aligned with the company's internal guidance on how the policy is to be enforced. Enforcement: Meta should notify Instagram users when a warning screen is applied to their content and provide the specific policy rationale for doing so.

Upheld cases

Case	Tigray Communication Affairs Bureau
Nature:	Violence and Incitement. Decision to remove a post threatening violence in the conflict in Ethiopia
Structure of Investigation	1. Following Meta's referral and the Board's decision to accept the case, the user was sent a message notifying them of the Board's review and providing them with an opportunity to submit a statement to the Board. The user did not submit a statement. 2. The Board asked Meta 20 questions. Meta answered 14 questions fully and six questions partially. The partial responses related to the company's approach to content moderation in armed conflict situations, imposing account restrictions for violations of content policies, and the cross-check process.
Public Comments:	The Oversight Board received and considered seven public comments related to this case. One of the comments was submitted from Asia Pacific and Oceania, three from Europe, one from Sub-Saharan African and two From the United States and Canada. The submissions covered the following themes: – the inconsistency of Meta's approach in the context of different armed conflicts; -the heightened risk accompanying credible threats of violence between parties during an armed conflict; – the problems with Meta's content moderation in Ethiopia and the role of social media in closed information environments;

– factual background to the conflict in Ethiopia, including the harm suffered by Tigrayan people and the role of hate speech against Tigrayans on Facebook in spreading violence;

– and the need to consider laws of armed conflict in devising policies for moderating speech during an armed conflict.

Oversight Board Decision	The Oversight Board upholds Meta's decision to remove the content for violating the Violence and Incitement Community Standard.
Policy Advisory Statement	Transparency: In line with the Board's recommendation in the "Former President Trump's Suspension" as reiterated in the "Sudan Graphic Video" Meta should publish information on its Crisis Policy Protocol. The Board will consider this recommendation implemented when information on the Crisis Policy Protocol is available in the Transparency Center, within six months of this decision being published, as a separate policy in the Transparency Center in addition to the Public Policy Forum slide deck. Enforcement: To improve enforcement of its content policies during periods of armed conflict, Meta should assess the feasibility of establishing a sustained internal mechanism that provides the expertise, capacity and coordination required to review and respond to content effectively for the duration of a conflict. The Board will consider this recommendation implemented when Meta provides an overview of the feasibility of a sustained internal mechanism to the Board.

Case	Sudan graphic video
Nature:	Violent and graphic content. Decision to restore a Facebook post depicting violence against a civilian in Sudan.
Structure of Investigation	1. Following Meta's referral and the Board's decision to accept the case, the user was sent a message notifying them of the Board's review and providing them with an opportunity to submit a statement to the Board. The user did not submit a statement. 2. The Board asked Meta 21 questions. Meta responded to 17 fully and four partially. The partial responses were to do with questions on measuring the impact of Meta's automated system on content on the platform and why the Violent and Graphic Content Community Standard does not contain a raising awareness exception.
Public Comments:	The Board received five public comments for this case. Two comments were from Europe, one from Sub-Saharan Africa, and two from the United States and Canada. The submissions covered the following themes: the need to adopt a more context-sensitive approach that would set a higher threshold for removal of content in regions subject to armed conflicts, so that less content is removed; the need to preserve materials for potential future investigations or to hold violators of human rights accountable; and that the newsworthiness allowance is likely to be applied in an ad hoc and contestable manner and that this practice should be reconsidered. In March 2022, as part of ongoing stakeholder engagement, the Board spoke with approximately 50 advocacy organization representatives and individuals working on reporting and documenting human rights abuses, academics researching ethics, human rights, and documentation, and stakeholders interested in engaging with the Board on issues arising from the Violent and Graphic Content Community Standard and its enforcement in crisis or protest contexts.
Oversight Board Decision	The Oversight Board upholds Meta's decision to leave up the content with a screen that restricts access to those over 18.
Policy Advisory Statement	Content policy: Meta should amend the Violent and Graphic Content Community Standard to allow videos of people or dead bodies when shared for the purpose of raising awareness of or documenting human rights abuses. This content should be allowed with a warning screen so that people are aware that content may be disturbing. The Board will consider this recommendation implemented when Meta updates the Community Standard. Enforcement: To ensure users understand the rules, Meta should notify users when it takes action on their content based on the newsworthiness allowance including the restoration of content or application of a warning screen.
Case	**South Africa slurs**
Nature:	Hate speech. Decision to remove a post discussing South African society under its Hate Speech Community Standard.
Structure of Investigation	1.Facebook notified the user that their post violated Facebook's Hate Speech Community Standard. Facebook stated that the notice to the user explained that this Standard prohibits, for example, hateful language, slurs, and claims about the coronavirus. 2.The user appealed the decision to Facebook, and, following a second review by a moderator, Facebook confirmed the post was violating. The user then submitted an appeal to the Oversight Board. 3.The Board asked Facebook how its market-specific slur list is enforced, and if a slur's appearance on any market list means it cannot be used globally. Facebook responded that its "prohibition against slurs is global, but the designation of slurs is market-specific, as Facebook recognizes that cultural and linguistic variations mean that words that are slurs in some places

	may not be in others." The Board reiterated its initial question. Facebook then responded "[i]f a term appears on a market slur list, the hate speech policy prohibits its use in that market.
Public Comments:	The Oversight Board received six public comments related to this case. Three of the comments were from Sub-Saharan Africa, specifically South Africa, one was from Middle East and North Africa, one was from Asia Pacific and Oceania, and one was from the United States and Canada. The Board received comments from stakeholders including academia and civil society organizations focusing on freedom of expression and hate speech in South Africa. The submissions covered themes including the analysis of the words "clever blacks," "n***er" and "k***ir;" whether the words "n***er" and "k***ir" qualify as hate speech; the user's and reporter's identity and its impact on how the post was perceived; and the applicability of Facebook's Hate Speech policy exceptions.
Oversight Board Decision	The Oversight Board upholds Facebook's decision to remove the content.
Policy Advisory Statement	Facebook should: Notify users of the specific rule within the Hate Speech Community Standard that has been violated in the language in which they use Facebook, as recommended in case decision 2020-003-FB-UA (Armenians in Azerbaijan) and case decision 2021-002-FB-UA (Depiction of Zwarte Piet). In this case, for example, the user should have been notified they violated the slurs prohibition.

Case	Alleged crimes in Raya Kobo
Nature:	Hate Speech. Decision to remove a post alleging the involvement of ethnic Tigrayan civilians in atrocities in Ethiopia's Amhara region.
Structure of Investigation	1. Meta notified the user that his post violated Facebook's Hate Speech Community Standard, but not the specific rule that was violated. The user then appealed the decision to Meta, and, following a second review by another moderator from the Amharic content review team, Meta confirmed that the post violated Facebook's policies. 2. The user then submitted an appeal to the Oversight Board. As a result of the Board selecting the case, Meta identified the post's removal as an "enforcement error" and restored it on August 27. Meta stated that it usually notifies users about content restoration on the same day. However, due to a human error, Meta informed this user of restoration on September 30.
Public Comments:	The Oversight Board received 23 public comments related to this case. Six of the comments were from Sub-Saharan Africa, specifically Ethiopia, one from the Middle East and North Africa, one was from Asia Pacific and Oceania, five were from Europe and ten were from the United States and Canada. The Board received comments from stakeholders including academia, private individuals and civil society organizations focusing on freedom of expression and hate speech in Ethiopia. The submissions covered themes including whether the content should stay on the platform, difficulties in distinguishing criticism of the TPLF from hate speech against the Tigrayan people, and Meta's lack of content moderators who speak Ethiopian languages.
Oversight Board Decision	The Oversight Board upholds Meta's original decision to remove the content. Given that Meta subsequently restored the content after the user's appeal to the Board, it must now remove the content once again from the platform.
Policy Advisory Statement	Meta should: rewrite Meta's value of "Safety" to reflect that online speech may pose risk to the physical security of persons and the right to life, in addition to the risks of intimidation, exclusion and silencing. Facebook's Community Standards should reflect that in the contexts of war and violent conflict, unverified rumors pose higher risk to the rights of life and security of persons. This should be reflected at all levels of the moderation process. Meta should commission an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumors that heighten the risk of violence in Ethiopia. The assessment should review the success of measures Meta took to prevent the misuse of its products and services in Ethiopia.

Looking at this table reveals that the appeal system mostly receives input from ourside sub-Saharan Africa. This raises the question of access to the appeal system by people from sub-Saharan Africa, particularly the subcommunities in the country of origin of the case who might be affected by the content.

OB decisions and national democratic governance and self-determination

To understand how the OB's decisions relate to democratic governance within the country of origin of the complaint it is necessary to look at the political systems in the concerned countries. This approach

can also help reveal how democracy at the national level functions in the context of hyper-globalization. As highlighted earlier, all five cases emanate from countries where either war or conflict is currently brewing, and/or where there is a history of racial violence and divides.

– Nigeria, ongoing political and religious tensions, embedded in the tribal, political, and religious fabric of that nation
– South Africa, racial and economic tensions dating to previous colonial governing systems
– Ethiopia has an ongoing regional/ethnic tension, brewing from political governing systems
– Sudan too has ethnic and political tensions stemming from issues of borders and contested leaderships shaped by colonial history

These historical and current political systems shape these society's values, sensitivities and norms, and these are often reflected in the behaviors of citizens, including in their engagements on online platforms. It is also worth noting that most of these countries are signatories to international human rights treaties and do have human rights laws and institutions. By appealing the decisions of the platform, users are indicating that content moderation does not show an understanding of the local context. In more direct terms, the platform's original decisions made based on its rules and guidelines appear to be one-dimensional. This is where the OB is intended to provide a broader perspective and consider local contexts. However, it appears that the decisions of the OB are also inadequate as they are made based primarily on a framing of the issues only based on universal values and laws even if the context requires deeper cultural/social and political sensitivity. This is demonstrated by the lack of direct use of local or regional laws in the cases at hand. By only paying attention to community standards and guidelines and the application of international human rights law, the OB reinforces the argument that platforms can undermine the self-determination of people within a national democratic space.

An alternative approach would have been for the OB to pay more attention to national legislative and human rights frameworks – and even regional Africa frameworks - and to convene more regional consultations. This would complement their existing process. By involving local and regional bodies including NGOs, regulatory bodies, and even digital oversight mechanisms at the African Union level and other relevant regional bodies they can make decisions that are more respectful of national democratic laws and processes.

The situation of applying universal laws and values to local contexts further translates to imposing global laws in the local context and disregarding local and regional guidelines (when in place) and laws and indirectly enforcing historical narratives of bigger power in local matters. In the cases under focus and informed by their review and decision-making processes, it is clear that the decisions of the OB did not consider or complement local laws and norms. The case of the Tigray Communication Affairs Bureau becomes a key example here. In this case, we see content created from within a specific region where people are disgruntled with the current political situation. Within that context, the content was flagged to be taken down and even though two Amharic-speaking reviewers "determined that the post did not violate Meta's policies and left it on the platform", Facebook through another layer of content review for "high-risk situations" went ahead and removed it. As demonstrated by the decision's explanatory notes, the Board did not consider how the local communications/editorial or media authority would have responded to or viewed this content. Consequently, this can be interpreted to be disrespectful of local contexts while amplifying international norms. By only looking at international human rights frameworks, the OB appears not to consider efforts within these countries to adhere to and implement the international human rights agreements to which they are signatories. This affirms the notion that the self-determination of nations/people is indeed difficult to achieve within hyper-globalization.

Covert ways of the enforcement of universal values and international laws within local contexts are demonstrated by the OB's investigation processes. They are carried out from afar and consultations are undertaken had limited public input from the affected country and the region. In studying the decisions of the OB in SSA, this study did not only find that the majority of the Sub-Saharan African cases received a low number of submissions from the public and third parties in general; it revealed that the majority of submissions were not from countries and regions that share history and political systems with the countries from which the cases came. More public and third-party submissions in support of the users from the relevant country and the region submitted to the OB before it made its final decisions might have presented the OB with a different viewpoint which it would have had to consider in its treatment of these cases.

For example, it might have encouraged the OB to put more effort into looking at the relationship between universal human rights standards and the dominant narratives of the countries of origin of the platforms. This relationship can be complex, made up of a mix of areas where these rights and standards conflict or coincide. This approach would respond to the challenges of current social media councils' regulations as criticized by Flew. et al (2020) that "the challenges of media regulation in the digital age arise from the complexity of regulating in a contested global arena where national policies are often in conflict and laws are not always enforceable in a straight-forward way".[10] In essence, considering national/regional policies and laws first, would add counter perspectives to this global perspective on local context as well as convey that the OB takes local contexts seriously, considering them before switching to applying global frameworks.

The question of self-regulation with regard to the knowledge structure of Meta and the OB remains central as posed by Mansell & Steinmueller (2020) who framed that self-regulatory practices run the risks of digital platforms making private decisions on behalf of the public. The OB makes 'public decisions' based on a public process, but it was set up as a self-regulatory arm of the former Facebook. I.o.w. Mansell's framing applies. The Board with funding from Meta engages consulting agencies for research aimed at informing the decisions they make on cases brought to them. The Board further sends questions to the platform to respond to (it does not always do so). The OB analyses the rules of the platform against the complaints of the users. While in some cases the Board seems to be referring to multistakeholder consultations which seem to have happened out of the scope of the cases in this regard (as in South Africa), it does not appear to have considered the media or social media regulations within a specific country. Ideally, this could have happened through consultations with the communications regulators within each country and or engagements with media councils, ombudspersons, and even the local internet governance systems in these countries.

In conclusion, this study finds that: the decision-making model of the Meta Oversight Board as a social media council is not entirely independent given that it omits to work with local policy, regulations, or other bodies, as well as taking global standards into account. It also fails to demonstrate how consultations at the local level are considered within their reviewing process. While it can be assumed that the consulting agencies they use would carry out this work, It is important for this fact to be reflected in their description of the process to avoid perceptions of being detached from local realities, applying instead globalized approaches to local contexts. Previous findings from research such as that by Barata

---

[10] T. Flew & M. Fiona &N. Suzor (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. Journal of Digital Media & Policy. 10. 33–50. 10.1386/jdmp.10.1.33_1.

(2020)11 support this conclusion by indicating that "the approach of the Oversight Board has exclusively been based on universal standards" which consequently disregards local standards and laws. While recognizing that the Oversight Board is overseeing decisions made by global platforms and must uphold international laws, this research has indicated that it has failed to apply what is widely understood as the multistakeholder approach to internet governance that uses consensus building in making decisions.

Recommendations

Given the findings of this paper, the following recommendations are proposed:

– The Oversight Board policies need to be explicit and intentional in giving consideration to national and regional laws and norms as the first step in dealing with cases under review
– The Oversight Board should engage directly with multistakeholder processes within the country and regions of their decisions as proposed by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression in 2019.
– The Oversight Board should indicate how and with who their research consultants engaged people and institutions (including media regulators and human rights institutions) in the relevant countries and reflect this within their case decisions.
– For the African Region, the Oversight Board should demonstrate its inclusion of engagements with regional laws and regional bodies as a necessity in trusting that their processes are focused on dealing with local contexts. This will ensure transparency and build the legitimacy of regional oversight structures and internet-related human rights norms and standards.
– The Oversight Board should invest heavily in awareness campaigns in Africa about their appeal procedures for cases as well as mechanisms of public input within their investigating processes.

---

11 J.Barata "The Decisions of the Oversight Board from the Perspective of International Human Rights Law" Special Collection of the case law on Freedom of Expression, Colombia University, available online at https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2022/10/The-Decisions-of-the-OSB-from-the-Perspective-of-Intl-Human-Rights-Law-Joan-Barata-.pdf

Elections and social media platforms in Kenya

Grace Githaiga
KENYA ICT ACTION NETWORK

Introduction

Elections are scheduled to take place in more than 70 nations in 2024, including some of the largest democracies on earth and others with weak or factually nonexistent democracies. A rising number of African countries with recent ones being Kenya in 2022, and Nigeria in 2023, have recently deployed tech for elections. This has been in the hope of inspiring confidence, increasing transparency and trust, especially in presidential elections that are mostly contested.

These recent elections have also seen a lot of use of social media platforms by politicians to share their manifestos, advance electoral ideals, conduct campaigns, motivate and galvanise their supporters around issues that are deemed of interest to these followers. Notably, in Kenya, some politicians attracted such huge following on such platforms as Meta and Twitter and ended up having ''broadcasting stations''[1] on these platforms, where they would have daily broadcasts targeting their supporters during the campaign season in 2022. In these daily broadcasts, citizens were sometimes exposed to fake news, disinformation, hate speech and fake photos of "large crowds" at campaign rallies.[2]

Candidates in Kenya frantically tried to catch the attention of the nation's 12 million social media users due to the growth in popularity of platforms like Twitter and Facebook.[3] A phenomenon occurred where requests to rent out Facebook pages and Twitter accounts from influencers with hundreds of thousands of followers became rampant, mainly for political purposes. These influencers gave politicians the chance to disseminate their views, address criticism, or even generate rumors about their rivals.[4] Very few candidates could therefore afford to ignore social media. This became good business for the influencers who would get paid a daily rate of 400 euros for a hashtag's development and maintenance.[5]

The two main presidential candidates, Dr. William Ruto of the Kenya Kwanza Alliance, and Mr. Raila Odinga of Azimio la Umoja-One Kenya Coalition Party "invested heavily in building digital war chests to interact with their supporters and market their manifestos and agenda".[6] Nonetheless, both coalitions suffered from fake news spear campaigns mostly fueled by influencers working for the different coalitions.[7]

---

[1] Kenya's President Ruto who has a 2.3M following would run daily broadcasts of his campaigns in 2022. https://www.facebook.com/williamsamoei.Another example is former Nairobi Governor Mike Sonko with a following of 2.5M https://www.facebook.com/search/top/?q=Mike%20Sonko and still runs Sonko TV which has 104K follower https://www.facebook.com/MikeSonkoTv001.

[2] Githaiga, Grace. 2022. Disinformation Factories in Kenya. https://www.linkedin.com/posts/gracegithaiga_election-influencers-for-hire-kenyas-disinformation-activity-6928056612398010368-VyzG/?trk=posts_directory.

[3] Caronline Rabut. 2022. Election influencers for hire: Kenya's disinformation factories. calendar.google.com/calendar/u/0/r?tab=mc

[4] AFP. 2022. Social media Influencers cash in as presidential election approaches in Kenya. https://www.africanews.com/2022/05/05/social-media-influencers-cash-in-as-presidential-election-approaches-in-kenya/.

[5] Ibid.

[6] Steve Omondi. 2022. How social media influencers shaped Kenya's 2022 General Election, https://mediainnovationnetwork.org/2022/08/29/how-social-media-influencers-shaped-kenyas-2022-general-election/.

[7] Ibid.

A study by the Mozilla Foundation, observed that social media sites such as Facebook, TikTok, and Twitter failed to uphold their promises of election integrity during Kenya's 2022 elections. Further, it was found that political advertising contributed to amplification of propaganda, and the platforms content labeling was unable to combat disinformation.[8] Twitter and Tiktok's labeling efforts were sporadic and ineffective in halting the spread of disinformation. The study further averred that this affected some political parties more negatively especially when these platforms went ahead to announce the election results ahead of the formal announcement, making them appear partisan. As for facebook, the platform was seen as lacking visible labels, which allowed for the dissemination of propaganda. The Mozilla study deduced that platforms therefore failed to live up to their promises to stakeholders to be reliable sources of election information and instead ended up being avenues where rumors, conspiracies, and false information thrived.

Kenya has a population of over 50 million, with half of them being under 35 years. Further, 12 million of them utilize social media.[9] Considering that a large part of the African electorate are young people who are on social media platforms, what ethical concerns does this raise for the platforms? How are platforms responding to these concerns considering some, such as facebook, twitter, tiktok have now become mainstream and especially during electioneering periods as observed in Kenya? What is the level of preparedness of these platforms in terms of content moderation?

This paper briefly responds to some of these issues drawing on Kenya's examples during the electioneering period of 2022.

Guiding questions

In this section, three guiding questions are addressed.

Ethical Concerns for Platforms

Social media platforms provided unprecedented opportunities to both candidates and their supporters to engage, during the electioneering period in Kenya. There was an increase in the number of politicians making use of social media platforms in 2022 compared to the 2017, and 2013 elections. But with this surge, the platforms were abused to amplify social differences and influence citizens' decision-making processes.[10] The social media space was transformed into a ferocious battleground where bloggers from Kenya's two main political parties spewed hate speech and used all kinds of dirty techniques to psychologically edge out their rivals in the hotly contested vote.[11] The platforms steadily accumulated virtual "litter" in the form of fake news, misinformation, disinformation, and propaganda from the pre-campaign period through the official three-month campaign period to the election and post-election period.[12]

---

[8] Odanga, Madung. 2022. Opaque and Overstretched: how platforms failed to curb misinformation during the Kenya 2022 election. https://foundation.mozilla.org/en/campaigns/opaque-and-overstretched-part-ii/#case-study-labeling-failures.

[9] AFP. 2022. Social media Influencers cash in as presidential election approaches in Kenya. https://www.africanews.com/2022/05/05/social-media-influencers-cash-in-as-presidential-election-approaches-in-kenya/.

[10] UNESCO. 2022. How UNESCO works to curb online hate speech and disinformation ahead of Kenyan general elections. https://www.unesco.org/en/articles/how-unesco-works-curb-online-hate-speech-and-disinformation-ahead-kenyan-general-elections.

[11] Steve Omondi. 2022. How social media influencers shaped Kenya's 2022 General Election. https://mediainnovationnetwork.org/2022/08/29/how-social-media-influencers-shaped-kenyas-2022-general-election/.

[12] Ibid.

There have been efforts by different stakeholders to confront the challenges. For example, the government's response has been through the establishment of the National Cohesion and Integration Commission (NCIC) in 2008.[13] This was done under the NCIC Act No.12 of 2008. The purpose of the NCIC was to create a national institution to advance national identity and values, reduce ethnic-political rivalry and violent crime, end racial, ethnic, and religious discrimination, and foster national healing and reconciliation. In application on social media, Section 62 of the Act outlaws speech that is designed to provoke contempt, hatred, hostility, violence, or discrimination against any individual, group, or community on the basis of ethnicity or race.[14]

Platforms such as Twitter, Meta, and TikTok all indicated that they were battling false information during 2022 Kenya's general elections.[15] They revealed that they had deployed moderators to remove harmful content. However, civil society groups felt that because of these moderators' lack of understanding of regional languages and dialects, and the lack of standards that take into consideration certain cultural and societal settings, bad content was spread swiftly with potentially harmful results.[16]

Several challenges still need attention from platforms. For example, there is the young population in Kenya on social media platforms and who are potential voters come the next general election. The platforms will need to engage this section of the population to raise awareness on responsible use of social media, but also collect their views on ways in which they think the platforms can respond to challenges for example of hate speech and disinformation. This can be done through focus groups, surveys, and in-depth interviews. These are methods for involving people in community development and asking for their perspectives.[17] Another challenge is that of the AI language tool(s) such as ChatGPT[18] that might see a significant amount of seemingly original content being produced by the use of generative language models, relieving propagandists of the need to repurpose the same text across several news websites or social media platforms.[19]

There is therefore a need for platforms to address these ethical concerns in a timely fashion, in order to contain the rise in political tensions using platforms, and especially during the electioneering period.

What are platforms doing– and is it enough?

There is no doubt that tech platforms continue to be used in Kenya to advance electoral ideals and manifestors. But with an increase in use, new challenges continue to emerge such as fake news, disinformation, propaganda, including insults against opponents, threats and incitement to violence,

---

[13] The National Cohesion and Integration Commission (NCIC). https://cohesion.or.ke/.

[14] RefWorld. 2013. Tackling Online Hate Speech in Kenya. https://www.refworld.org/docid/51308a312.html.

[15] Nita Bhalla. 2022. Call to action: Kenyan elections under threat of hate speech. https://www.csmonitor.com/World/Africa/2022/0627/Call-to-action-Kenyan-elections-under-threat-of-hate-speech.

[16] Ibid.

[17] Behaviorial Insights Team. 2022. How can citizens shape social media platforms? https://www.bi.team/blogs/how-can-citizens-shape-the-future-of-social-media-platforms/.

[18] Start using Chat GPT. https://ai-pro.org/start-chat-gpt/?ppg=03&keyword=chatbot%20ai&adid=647798462831&gclid=Cj0KCQiAgaGgBhC8ARIsAAAyLfF6TOy3RWazp-F3Go_Z8mgNKBRQQGhoUzUfqFdQ8X5VVmR1RbaGeQcaAqU1EALw_wcB

[19] Timcke, Scott . 2023. Generative Language Models in Algorithmic Social Life: Some Concepts and Considerations. Research ICT Africa. https://researchictafrica.net/2023/01/12/generative-language-models-in-algorithmic-social-life-some-concepts-and-considerations/.

defamatory language, propaganda, and covert hate speech etc. Such challenges are contributing to the calls from different stakeholders for content moderation on platforms.

In July 2022, the Kenyan government through the NCIC, indicated that Meta (Facebook) had neglected to fix the problem of hate speech ahead of August's general elections. Using a report produced by Global Witness, there was an indication that English and Swahili, the two official languages of Kenya, were not found to contain hate speech adverts on Facebook.[20] This was not the situation and the issue was that Facebook failed to flag hate speech adverts in either language. Accordingly, the platform was given one week to adhere to laws pertaining to hate speech.[21]

Facebook was not the only culprit. According to research by the Mozilla Foundation, TikTok was seen as a "platform for quick and far-reaching political disinformation".[22] In fact, "[e]ven videos that appear to be in clear violation of TikTok policies seem to have been amplified by the algorithms".[23] In response, Tik Tok removed the videos flagged by the Mozilla Foundation, which the platform then determined to be against its community guidelines.

Twitter indicated that it had employed skilled moderators and worked with independent fact-checkers as part of their aggressive disinformation-fighting strategy.[24] Further, the platform averred its commitment "to protect the health of the electoral conversation." However, it has been felt that the bringing down of offensive posts sometimes takes too long to prevent damage[25].

Engagement of civil society

As part of its civic duty and election work, KICTANet[26] a multistakeholder think tank for ICT policy and regulation and a catalyst for reform in the ICT sector, convened dedicated engagements for civil society actors with both Meta and Twitter. The Network also participated in a meeting that Tik Tok convened for civil society organisations.[27] The dedicated engagements between KICTANet and platforms discoursed on concerns about social media use during elections. It was disclosed that as a way of responding to concerns that had emerged during elections, the three platforms had put in place various mechanisms which aimed at monitoring and addressing false information and hate speech on their platforms. The mechanisms included, among other things, encouraging fact-checking, promoting instructional materials on false information and hate speech, and providing links and election-related information. However, stakeholders felt that more could be done in combating the challenges identified, in order to inspire trust and confidence in users of these platforms. For example, it was expressed that

---

[20] Samuel Gebre. 2022. Social media platforms under scrutiny ahead of Kenyan elections, https://techxplore.com/news/2022-08-social-media-platforms-scrutiny-kenyan.html.

[21] Ibid.

[22] Ibid.

[23] Andrew Deck. 2022. Hate speech and disinformation spike on TikTok in run-up to Kenya's elections.

https://restofworld.org/2022/hate-speech-and-disinformation-spike-on-tiktok-in-kenya-election-run-up/

[24] Nita Bhalla. 2022. Online disinformation stokes tensions as Kenya elections near. https://www.context.news/digital-rights/online-disinformation-stokes-tensions-as-kenya-elections-near.

[25] My own conclusion after conversing informally with colleagues.

[26] www.kictanet.or.ke. KICTANet's guiding philosophy encourages synergies for ICT policy-related activities and initiatives. The network acts as a catalyst for reform in the ICT sector and is guided by four pillars: policy advocacy, stakeholder engagement, capacity building, and research.

[27] Mwendwa Kivuva. 2022. Emerging Concerns on Social Media use in the upcoming 2022 Elections.

2022, https://www.kictanet.or.ke/emerging-concerns-on-social-media-use-in-the-upcoming-2022-elections/.

the platforms needed to make enough of an investment or take enough action to identify, prohibit, and prevent hate speech, false information, and disinformation on their platforms especially during the electioneering period.[28]

Level of preparedness for content moderation

Social media platforms have found themselves under scrutiny, in particular on how they moderate content during elections. Their level of preparedness has been questioned in particular when it comes to elections in African countries. There have been questions around the criteria they use to moderate content, the tools they use, especially where there exists a multiplicity of local languages as well as dialects.

Organizations monitoring social media and fighting to defend election integrity, noted that social media corporations invest less and provide fewer tools in the Global South than in the North.[29]

There is also now the complex language processing tools like ChatGPT which may have an impact on the operations on social media.[30] It is touted that language models might be able to compete with human-written content for a reasonable price. As such, any potent technology might provide specific advantages to propagandists who choose to use them.[31] It remains to be seen how the platforms will handle these new AI language tools, especially in African elections and whether they will be used to create automated propaganda. It will be interesting to see how the tool for example will write local languages and dialects and the slant in messaging.

Conclusions

Platforms are going to increasingly be used during elections in Africa, mainly as a double sword: to campaign and share political parties' manifestos, sow dissent and spread propaganda of opponents.

In light of this and considering the need to have content moderation policies that might at times need to be context specific, this paper notes that every social media firm has its own policies and procedures regarding online harm, such as with varying definitions of terms like misinformation and disinformation. As such, it would be useful to consider if there is a need to put pressure on social media platforms to adopt uniform definitions for all online lies and disinformation.

There is a lack of clarity on which languages are moderated by the platforms in the different African countries. It would be important to have an idea how this is determined, since Africa has many languages that have different dialects and moderators would need to understand the nuances.

This paper notes that platforms endeavored to engage with stakeholders in Kenya during the electioneering period. This was commendable as engagement with key stakeholders brings about an understanding of what can be done and what is not feasible. Stakeholders, and by and large the public get to learn of the platform's community standards, and that there are buttons for reporting hate, misinformation and other social media bad behaviour. Most importantly, ways need to be developed in

---

[28] Ibid.

[29] Digital Action. 2022. Roundtable briefing: 2024: Global Year of Democracy.

[30] David Silverberg. 2023. Could AI swamp social media with fake accounts? https://www.bbc.com/news/business-64464140.

[31] Goldstein, Josh A et. al. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://www.arxiv-vanity.com/papers/2301.04246/.

which stakeholders can contribute ideas to some of the challenges raised in a timely fashion. Platforms can receive a larger range of opinions from the user community by establishing collaborative, deliberative spaces. These can be through townhalls, and quarterly engagements akin to the convenings by KICTANet in 2022 in which platforms got to hear views from civil society actors.

Policy Recommendations

This paper makes the following recommendations to several stakeholders as follows:

Platforms

– In the context of the global election cycle in 2024 and considering that platforms have become sources of exchange of election information and disinformation, there is a need for preparation before the election years and periods of different countries. They need to put in place mechanisms for the challenges provided by the digital dangers on their platforms early enough and not necessarily during the election year. The mechanisms would include for example policies on how to deal with hate expressed in local languages, what to do with verified accounts of sitting leaders in the country if they are used for disinformation or hate speech, etc.

– Platforms must be deliberate in engaging with key stakeholders to inform them of what options are available for reporting electoral offenses and encourage them to give their views in order to strengthen the platforms. The engagements need to be regular and not confined to when elections are a few months away as there might not be adequate time to respond to new ways of doing things, or effecting changes recommended.

– Platforms will need to be intentional and create more awareness to let citizens know that they too have an option to participate, contribute and have their views taken on board. Communication to stakeholders should be simplified in such forms as animations, short videos, and in a way that inspires confidence in ordinary citizens (example here). In addition, there must be a way to demonstrate that citizens' opinions have been taken on board through clear feedback mechanisms, and not just taken through a participatory process as a mere ticking of boxes.

– Social media corporations, through their content moderation advisory groups, will need to deliberately consider users as citizens and collaborators in determining the direction in which the platforms determine content moderation models. The advisory groups' decisions need to have tangible effects, as currently, there is very little information on what these advisory or councils are doing or have been able to achieve.

– It is important for platforms to to be in tune with public values in order to respond accordingly.

Civil Society

– Future elections could see an even greater flood of false information. Some of these platforms such as Meta have an Eastern Africa office based in Nairobi, while others have no local presence, with their Africa policy teams based in Western Countries. Civil society therefore must be deliberate and strategic in engaging these platforms and point out local challenges. Further, they will need to make suggestions on safeguards, especially on the need for platforms to be accountable. In addition, they can also provide proposals on how context specific hate and disinformation on platforms can be minimized by respecting the local values that embrace safety and confidence of users wherever they are.

Donors

– There is a need for international development partners, donors and funders to avail resources to Global South Civil Society organisations in an effort to support their work in campaigns of engagement and calling out platforms on their lack of commitment to having safeguards and measures to combat disinformation and electoral lies on their platforms. Civil society organisations can push for the need to have policies in particular as they relate to the Global South when it comes to electioneering time. But this can only be achieved if resources are available to engage in meaningful and tangible work before, during and after a general election.

A Proposal on Voluntary Social Media Councils

Kuda Hove
INDEPENDENT RESEARCHER, SOUTH AFRICA

Introduction

Social media platforms have changed the way people connect with each other. Platforms such as *Facebook, Twitter, Reddit,* and *LinkedIn* have billions of users across the world including Zimbabwe. One thing that most mainstream social media platforms have in common is that they are developed and owned by entities based in the United States of America, Europe, or China in the case of TikTok. A platform's country of origin influences that platform's governance structure as well as the platform's policies. This is problematic given that these platforms are global, operating across a range of varied cultural and political environments and/or contexts.

Content moderation decisions often show how challenging it is to come up with one central content moderation policy which applies to a global range of users. For example, Facebook and Instagram initially absolutely barred the sharing of images and videos of topless women. This made sense from a Western perspective but in places like the Zulu nation in South Africa or in eSwatini, it is normal for girls and young women to take part in public cultural events topless. Sharing images of such cultural events was not possible as it violated the platform's community guidelines. In this instance, Facebook's community guidelines restrict Zulus and the Swati from exercising their full cultural expression. However, with time these community guidelines have evolved to include exceptions in which nudity is allowed under specific circumstances.

This is just one example of many where platform governance falls short of giving users, especially those outside of the US and Europe the ability to freely express themselves in a legitimate manner. Related to this, the global nature of platforms means that they are not easy to regulate or hold accountable when they contribute to the violation of fundamental rights in a jurisdiction such as Zimbabwe. This brief article proposes a form of social media platform governance that is suitable to protect the rights of Zimbabwe based users without unjustifiably restricting the enjoyment of their fundamental rights.

Can we learn from regulation of print media?

In Zimbabwe and other economically similar African countries, social media and instant messaging platforms have become sources of news and information seen by some as being of similar value as traditional print media sources. In a country with high inflation rates and low disposable income, most Zimbabweans cannot afford to buy print media in the form of magazines and newspapers. WhatsApp and TikTok based news sources and Facebook based newsgroups have mushroomed to inform audiences which otherwise would have been informed by traditional print media.

These online based platforms unfortunately can also be used to spread misinformation, manipulate voters during the run-up to elections or a key referendum, and even to fan genocide. This is why there is a need to moderate the content shared over social media platforms and also to put in place some form of regulatory frameworks to safeguard service users in a given jurisdiction.

To mitigate some of the harms which may be caused by the use of social media platforms, there has been a call for the introduction of effective oversight measures to ensure that content shared over those platforms does not amount to misinformation, hate speech or other forms of speech which may incite or

lead to the violation of certain fundamental rights. Social media platforms have, for example, come up with content moderation policies but one main shortfall and criticism of content moderation policies is that they are not flexible enough to apply fairly across the different cultural, political, and racial contexts within which one social media platform, such as Facebook operates. Some social media companies have also used their resources to create what may be referred to as social media councils. But these social media councils face the same problem as the platforms and their content moderation policies – they are not flexible enough to adequately apply to all global contexts.

It is my submission that similarities between print media and social media platforms mean that some regulatory print models for the print media can be modified and applied to social media platforms. An interesting prospect is the voluntary media council method as seen in the operations of the Voluntary Media Council of Zimbabwe (VMCZ).

 The VMCZ is an independent, non-profit, self-regulatory body that promotes ethical and professional standards in Zimbabwe's media industry. The VMCZ was established by a group of media stakeholders who recognized the need for a self-regulatory mechanism to improve the quality of journalism, promote media freedom and more importantly, promote media accountability in the country. The VMCZ aims to promote a culture of ethical and professional journalism in Zimbabwe by providing a platform for media professionals and media consumers to engage in constructive dialogue on media ethics and best practices. Some of the country's largest independent media groups have volunteered to be bound by the VMCZ's Code of Conduct.

The VMCZ also seeks to promote media accountability through its complaints mechanism which members of the public can use to lodge complaints against media organisations or journalists who breach the VMCZ Code of Conduct they volunteered to be bound by. The VMCZ is not a regulatory authority and does not have the power to sanction or punish media organisations or journalists who violate ethical or professional standards. But the organisation does facilitate dialogue between the aggrieved party and the media organisation accused of breaching the VMCZ Code of Conduct.

VMCZ has managed to get some media journalists and media houses to print apologies and retract stories where they got the facts wrong or where the reporting breaches VMCZ's Code of Conduct. At a national governance level, VMCZ has been useful in lobbying the government for an independent Zimbabwe Media Commission.

What would a voluntary social media council look like?

Social media councils are often set up to perform an oversight role to evaluate the fairness of content moderation decisions taken by the social media platform they relate to. Statistics from Facebook's Oversight Board show that only 5 of the 46 complaints finalised between December 2020 and March 2023, were from Sub-Saharan Africa – specifically from Ethiopia, Sudan, Nigeria, and South Africa. This shows that a number of violations which take place on social media platforms go unreported if those violations take place in jurisdictions like Zimbabwe.

This could indicate that users are not aware of appeal processes which in turn may reflect a lack of awareness about the existence of social media councils like the Oversight Board. Setting up social media councils at national level could be useful in informing users about complaints and appeals procedures available to them. Additionally, establishing national level social media councils ensures that any local contexts are considered when investigating and handling a complaint.

The biggest challenge to the establishment of national level social media councils is the imbalance of power between most of the social media platforms and small economic markets like Zimbabwe, a low-income country. Social media platforms do not prioritise lower income countries because of the comparatively low return on investment they offer. High internet access costs in Zimbabwe mean that the number of users is comparatively lower than in bigger economies where people have more disposable income and enjoy higher levels of access to the internet.

Zimbabwe and other similar countries have no means to establish jurisdiction over foreign based social media platforms with financial reserves that are several times bigger than Zimbabwe's average national budget and GDP. As such, national social media councils are likely to be voluntary in nature with social media platforms choosing or volunteering to participate in these national media councils. Participation does not require that social media platforms establish a national office in each country, their participation can be virtual. In this way, content moderation teams can work with different national social media councils in each country to decide whether planned content moderation decisions are correct under the given circumstances.

Founding stakeholders for each national social media council stakeholders could be a mix of civil society actors, academics, users and representatives of special interest groups and minorities. The key is that these various stakeholders understand and have expertise in matters relating to free expression, access to information and related rights. It may also be useful to have public entities drawn from organisations such as National Data Protection Authorities, Free Speech Ombudsman and Constitutional Commissions tasked with promoting the enjoyment of the right to free expression and the related information rights.

A user driven regulatory approach is possible, as illustrated to a certain extent by Reddit's current content moderation process. Reddit's Content Policy is made up of eight rules which are universally applicable and stated in simple to understand language. These rules are enforced across each Reddit community not by a team of content moderators sitting in a regional Reddit office, but each community is moderated by its own community members who have been granted moderator status. These community level moderators have to abide by the terms of Reddit's Moderator Code of Conduct. This code of conduct provides oversight mechanisms for community moderators' activities.

How sustainable is this model?

The operational costs of these national social media councils would be minimal especially in instances when they are held virtually. Where possible, social media platforms could also support the maintenance of national social media councils since their work contributes to a safer user environment for each platform's users. Governments can also be instrumental in promoting the work that such national social media councils do. Research labs and civil society organisations can in a similar way support national social media council initiatives as a way of promoting their mission to protect free expression and information rights.

In instances when social media platforms do not voluntarily join national social media councils, these councils can play an advisory and referral role to global media councils such as Facebook's Oversight Board. In this way a Zimbabwean social media council would help users report violations of their privacy and refer those violations to the Oversight Board for redress. Additionally, national voluntary social media councils can also come together at regional levels to increase their chances of attracting validation/ the attention of social media platforms.

In conclusion, national social media councils, if executed well, can provide a legitimately independent, participatory, and transparent way to ensure user participation in content moderation and the regulation of social media platforms in each country. These social media councils would also be helpful in promoting awareness of safety measures and complaints processes across the various social media platforms.

Building and strengthening rights-based social media platform governance in Africa through national human rights institutions

Tomiwa Ilori
CENTRE FOR HUMAN RIGHTS, UNIVERSITY OF PRETORIA, PRETORIA, SOUTH AFRICA

Introduction

Most social media platform governance approaches in African countries are fraught with lack of trust and legitimacy. This lack is often due to the enormous rule-making and decision-making powers wielded by both African governments and social media platforms with respect to what stays on platforms and what does not. For example, the laws, policies and processes made by African governments to regulate online harms are often at variance with international human rights standards while social media platforms are distant from the contextual realities required to regulate online harms in African countries.

As a result of these, online rights are often violated at will while online harms continue to grow at an exponential rate. These violations and the growing threat of online harms have therefore made it necessary to rethink our idea of 'stakeholders' involved in the regulation of online harms beyond traditional government actors, social media platforms, and civil society in African countries.

Missing from most ideas about social media platform governance in African countries, especially as it concerns protecting online rights and preventing online harms, is the legal mandate of national human rights institutions (NHRIs). Therefore, this contribution examines the roles African NHRIs can play to build and strengthen rights-based social media platform governance in African countries.

A brief overview of African NHRIs

NHRIs are unique statutory domestic institutions with a constitutional and/or legal mandate to protect and promote human rights in national contexts. According to Sundström, there are four types of NHRIs in African countries. They are:

(1) classical ombudsmen (COs), with a mandate focused on maladministration, (2) human rights commissions (HRCs), with a broad human rights mandate, (3) human rights ombudsmen (HROs), with a mandate of both maladministration and human rights, and (4) specialised institutions, such as children's ombudsmen, with a narrower mandate.

Types of African NHRIs
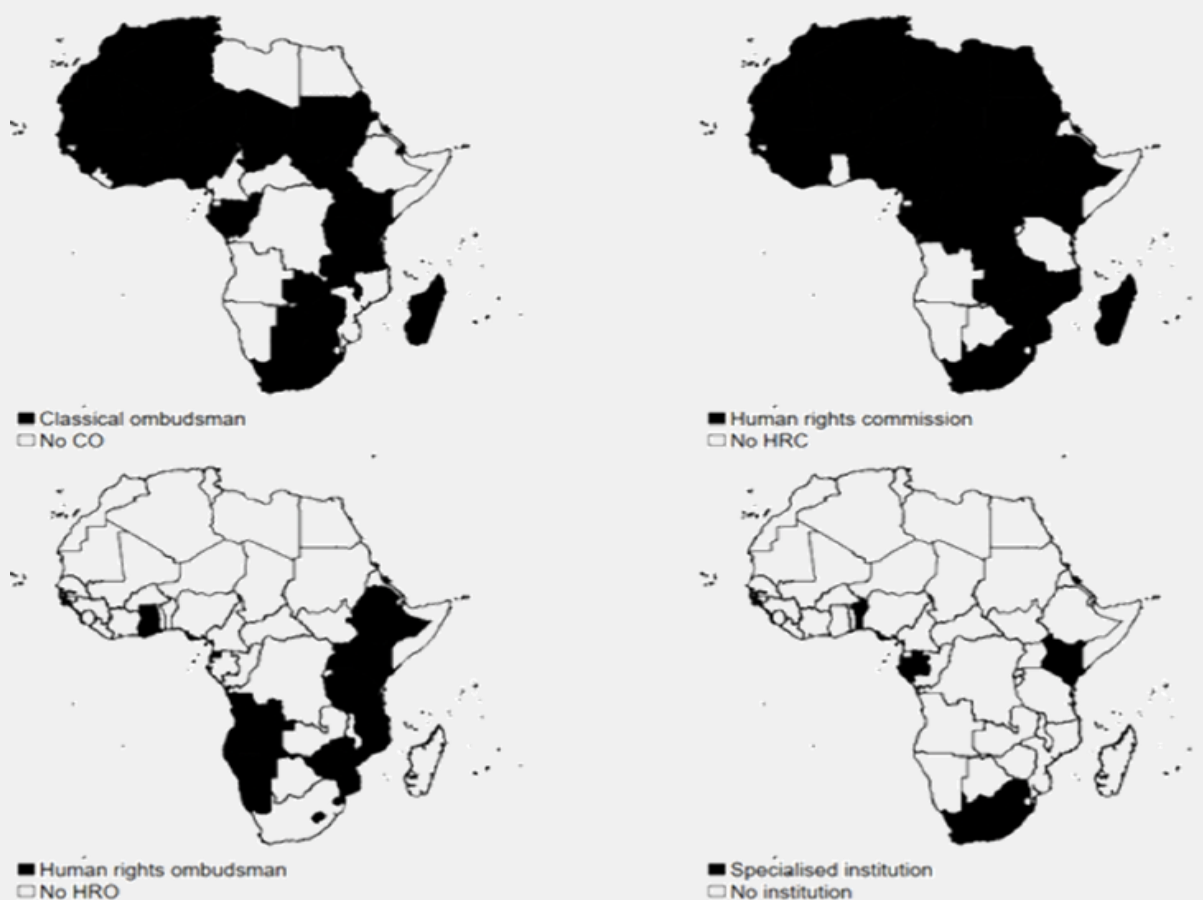
Source: Sundström, 2022

The first African NHRI was a classical ombudsman established in Tanzania in 1966. Currently, there are at least 47 NHRIs in African countries. Thirty-three out of the 54 African NHRIs have more than one type of NHRI. Most African NHRIs have both protective and promotional mandates and these include but are not limited to monitoring and reporting the human rights situation; providing advice to governments and others; delivering human rights education programmes; cooperating at the national level with key partners; and engaging with the international human rights system.

With respect to the last (but not the least) mandate, NHRIs have a key responsibility to ensure that key international human rights instruments are considered and implemented in local contexts. One of such key instruments is the United Nations Guiding Principles for Business and Human Rights (UNGPs) which has been further applied to states' obligations in ensuring rights-based social media platform governance in their various contexts.

Directed to social media companies, the human rights principles to be applied include ensuring human rights by default through terms of service, clear and determinable moderation rules, bases and extent of content moderation decisions, non-discrimination of protected characteristics, prevention and mitigation, transparency, and many others. Given NHRIs' mandates, they are well-positioned to monitor the implementation of these principles in domestic contexts by providing an anchorage for rights-based policies. One of such policies could be a soft law instrument such as a charter that sets out roles and

responsibilities of NHRIs and other actors with respect to the above-mentioned principles. However, despite having such a mandate and potential, NHRIs are faced with several challenges when it comes to social media platform governance in African countries.

Challenges facing NHRIs and social media platform governance in Africa

Social media platform governance in African countries is often seen from the vantage point of preventing online harms and protecting human rights. However, social media platform governance also includes economic and legal dimensions such as taxation, labour, competition and many others. But, so far as it concerns social media platform governance in African countries, human rights protection should rightly take centre stage for two major reasons. One, as critical stakeholders, many African governments have egregious internet freedom records, and two, power imbalances exist between states and social media platforms on the one hand and online publics on the other.

These reasons require us to think creatively about not just providing norms for social media platform governance but also think through how these provisions can be contextually relevant and effective domestically in African countries. In this regard, NHRIs are not only strategic domestic actors with respect to their main mandate to ensure national protection and promotion of human rights, they also serve as 'bridges' for mainstreaming international human rights standards such as those on social media platform governance into national contexts. Therefore, when it comes to the aspect of social media platform governance that involves human rights protection or mechanisms, including charters and social media councils, NHRIs need to play crucial roles.

However, African NHRIs face a number of challenges that could make playing these roles difficult. For example, as of November 2022, all the 28 African NHRIs rated based on the Paris Principles (Principles Relating to the Status of National Human Rights Institutions), which sets the minimum credible and operational standards that NHRIs must meet, received an 'A' status rating. This meant they were fully compliant with the Principles. However, out of this number, only four NHRIs including Ethiopia's Human Rights Commission, South Africa's Human Rights Commission (SAHRC), Nigeria's National Human Rights Commission (NHRC), and Kenya's National Commission on Human Rights (KNCHR) touch on the relationship between NHRIs and online rights protection as an objective in their annual strategic plans.

At the regional level, the Network of African National Human Rights Institutions' (NANHRI) Strategic Plan (2021-2025) highlights 'emerging and evolving issues' as some of the key strategic challenges facing NANHRI but does not expressly mention online rights protection. These challenges point to NHRI's lack of readiness to provide the capacity necessary for social media platform governance in African countries. In addition to this, NHRIs also have other capacity challenges related to financing, leadership, composition, human rights literacy and others.

Despite the challenges, it is not all gloomy. While, in non-African NHRIs have embraced their mandates in the facilitation of human rights protection in the context of data-driven businesses and recommendations for an Artificial Intelligence Safety Commissioner, African NHRIs such as the SAHRC and the NHRC are also seeking to exercise their mandate in the area. For example, SHRC is due to release a social media charter (SMCh) in South Africa, while NHRC is the main implementation actor in a proposed law on digital rights and freedoms in Nigeria. However, these taken together still point to a weak relationship between social media platform governance and NHRIs in Africa, as, in the larger context, it is almost non-existent, and what this points to, is the need to build and strengthen such a relationship.

Building and strengthening a rights-based social media platform governance through African NHRIs

The European version of NANHRI, the <u>European Network of Human Rights Institutions (ENNHRI)</u>, has highlighted the debates in online content regulation and identified four main ways through which NHRIs can protect online rights and freedoms including providing advice, advocacy, implementing the Marrakech Declaration into the expanding civic space, and promoting and protecting human rights defenders, and monitoring and reporting in online and offline civic spaces. This suggests that not only should NHRIs build their domestic capacity, they should also work together as a strong unit at the regional level.

At the domestic level, efforts by NHRIs in South Africa and Nigeria point to norm-setting and decision-making with respect to social media regulation and broader digital rights issues. While such efforts are crucial, in addition to them, NHRIs need to also get actively involved in playing an important role to ensure a rights-based approach to social media platform governance in African countries. This way, NHRIs lead on building and strengthening the <u>capacity</u> of duty-bearers to meet their obligations while also encouraging rights holders to claim their rights. Some of the ways through which such a role can be played include strategic collaboration, policy advocacy, capacity building and radical participation.

Strategic collaboration: Granted that social media platform governance is a complex subject involving diverse actors and interests, such a complex subject can be brought apart and pieced together especially at the domestic level by NHRIs and other major stakeholders. This is because NHRIs are strategic actors with respect to domestic implementation of human rights and this strategic positioning also places them amidst other critical actors such as governments, regional institutions, international organisations, academia, civil society and others.

Such collaborations should provide a multi-stakeholder platform for addressing some of the issues addressed above. For example, NHRIs, in partnership with these stakeholders may publish and disseminate a guide for NHRIs on the regulation of digital technologies in Africa. The purpose of the guide would be to provide for the responsibilities of NHRIs in the regulation of digital technologies and also provide specifically for the role of NHRIs in social media platform governance in African countries.

Policy advocacy: NHRIs need to include policy advocacy for digital rights protection, including those for social media charters in their strategy documents. Such advocacy may include research, workshops, seminars with strategic stakeholders referred to above on how to embark on a digital rights legal reform in African countries. This would provide for adequate planning, execution and measurement of specific activities towards such regulation. Another means of carrying out such advocacy is through legal advocacy. Such advocacy will involve providing advice on the human rights impacts of proposed digital policy laws, and where necessary, support public interest litigation.

In addition to this, new social media regulatory mechanisms such as <u>social media councils</u> (SMCs) are now being proposed by international actors. SMCs are independent, multi-stakeholder ombudsman mechanisms modelled after press councils that enable industry-wide complaint mechanisms and promote remedies from violations. Seeing the attempt by the SAHRC to create a SMCh which is usually the basis for establishing SMCs and based on Sundström's typology of NHRIs, SMCs can be established as classical ombudsperson, human rights commissions, human rights ombudsperson or specialised institutions. However, any preference out of these types must be thoroughly critiqued to justify its adoption.

Capacity building: Capacity building in this context refers to equipping NHRIs with the ability to perform their roles in social media platform governance in African countries. This includes but is not limited to building and strengthening NHRIs' educational and financial capacity. With respect to educational capacity building, there is a need for more knowledge exchange among NHRIs and other stakeholders about the opportunities and challenges with respect to social media regulation in their countries.

Such knowledge exchange can be carried out through workshops, seminars and fora. With respect to financial capacity, African NHRIs should expand their strategic partnerships in order to source for more funds to address their financial constraints. This is because one of the means of guaranteeing the independence of public institutions such as NHRIs is their financial sustainability and independence. However, in order to safeguard this independence, NHRIs should develop and publish an ethical funding policy.

Radical participation: One of the biggest challenges facing democratic social media regulation across the world today is how to make it radically participatory. Such radical participation involves giving everyone opportunities to influence how social media platforms are governed. This radical participation is also becoming more necessary given the public power wielded by social media companies.

This is a mammoth task not only because of the sheer legal and administrative measures required to facilitate such participation but also because providing such radicality requires grassroots participation. While NHRIs do not have answers to all the problems of social media regulation in African countries, they are strategically placed to implement these measures and ensure such participation especially at the domestic level. One of the measures that could drive such participation are SMCs. However, whether they will ensure this kind of radical participation in African countries is still left to be seen.

Conclusion

In order to start the serious conversations about social media platform governance in Africa, we need to work with what we have. And with this in mind, rights-based content regulation cannot keep playing catch up with ever-evolving social media platforms. Therefore, we will need to devise creative means for domestic participation in social media platform governance – and NHRIs provide such means. Prior to the 2000s, data protection developments in Africa were limited. Building on constitutional provisions on the right to privacy as what we have, today, not only do 35 African countries have data protection laws, at least 18 of them now have unique institutional mechanisms -- data protection authorities. This is still very far from what a strong regional data protection landscape should look like, however, they are positive developments. With South Africa's SAHRC set to release a Social Media Charter, African countries are on another journey of what is possible with what they have -- NHRIs -- and it is high time they took a lead as strategic stakeholders in building and strengthening rights-based social media platform governance in Africa. This includes NRHIs building and strengthening strategic collaboration, policy advocacy, capacity building and radical participation.

Unpacking the Tensions: A Comparative Analysis of DNS Abuse Mitigation and its Impact on Business and Human Rights

Ephraim Percy Kenyanito
QUEEN MARY UNIVERSITY OF LONDON, LONDON, UNITED KINGDOM

Introduction

In today's digital age, the internet has become an integral part of our daily lives. With the ever-increasing use of the internet, it has become a powerful tool for communication and expression of thoughts and ideas. However, the unrestricted use of the internet has also brought forth numerous challenges, including DNS abuse, intermediary liability, and censorship, which affect the fundamental right of freedom of expression.

DNS abuse refers to the malicious use of the Domain Name System (DNS), which is responsible for translating human-readable domain names into IP addresses. DNS abuse includes activities such as phishing, malware distribution, spamming, and other malicious activities that exploit the DNS system. We note that DNS Abuse cannot be solved without imposing Intermediary Liability on DNS operators. Intermediary liability, on the other hand, refers to the legal responsibility of intermediaries, such as internet service providers (ISPs), for the content posted on their platforms. Finally, censorship involves the control of information flow on the internet, with the aim of suppressing or controlling certain types of content.

In three previous blogs, this author has highlighted the risks of regulating DNS Abuse without clear definitions and without respecting human rights.[1]

In this article, we will explore the complex issues surrounding the use of the internet, including DNS abuse, intermediary liability, and censorship, with a focus on the institutional, legal, and contractual frameworks in place to mitigate these challenges. We will also explore the impact of these challenges on the fundamental human right of freedom of expression.

Institutional Framework

The internet is a decentralized network of networks that operate under the stewardship of various organizations, including the Internet Corporation for Assigned Names and Numbers (ICANN). ICANN is responsible for managing the DNS system and ensuring that it operates smoothly and securely. The ecosystem surrounding ICANN is vast, comprising various stakeholders, including registries, registrars, and internet service providers (ISPs).

The role of these stakeholders is crucial in the implementation of DNS abuse mitigation measures and accruing liability. Registries are responsible for the management of top-level domains (TLDs), such as

---

[1] https://www.article19.org/resources/online-freedoms-safeguards-must-be-balanced-with-free-expression,
https://www.article19.org/resources/icann-content-moderation-at-the-infrastructure-level-is-a-dangerous-move,
https://www.article19.org/resources/europe-content-moderation-at-infrastructure-level-must-respect-human-rights.

.com, .org, and .net, while registrars act as intermediaries between registries and domain owners. ISPs provide access to the internet and are therefore responsible for ensuring that their networks are secure and free from DNS abuse.

To address DNS abuse, ICANN has implemented several policies and programs. One of these is the Contractual Compliance Program, which enforces the terms and conditions of ICANN's agreements with domain name registrars and registries. This program ensures that these entities comply with ICANN's rules and regulations related to DNS abuse.

ICANN also has a Global Domains Division (GDD) that works to protect the security and stability of the DNS. The GDD collaborates with other organizations, such as law enforcement agencies, to identify and mitigate DNS abuse. In addition, ICANN participates in industry groups and initiatives, such as the Anti-Phishing Working Group (APWG) and the Messaging, Malware, and Mobile Anti-Abuse Working Group (M3AAWG), to address DNS abuse.

In addition to ICANN, other organizations play a vital role in the management of the internet, including the Internet Engineering Task Force (IETF), the World Wide Web Consortium (W3C), and the Internet Governance Forum (IGF). The IETF develops and promotes internet standards, while the W3C develops web standards. The IETF has also taken steps to address DNS abuse. For example, the DNS Abuse Working Group, a subgroup of the IETF's Security Area, focuses on developing technical solutions to address DNS abuse. The working group has published several documents related to DNS abuse, including RFC 7706, which describes a method for identifying and mitigating DNS-based threats. The IGF, on the other hand, is an open forum for stakeholders to discuss internet governance issues.

Legal and Contractual Framework

Registry policies play a critical role in mitigating DNS abuse, intermediary liability, and censorship. The policies of select registries such as (.org[2]), (.ke)[3], (.cn)[4], (.eu)[5], (.br)[6], (.mw)[7], and (.tz)[8] have been analyzed against international human rights law and best practice.

The .org registry has implemented a domain abuse activity reporting system, which aligns with best practice for mitigating DNS abuse. The .ke registry is in the process of developing a draft policy on the prevention and mitigation of DNS abuse, which the author intends to analyse in future publications for its consistency with international human rights law. However we can note that previous actions at the registry indicated grounds for concern.[9]

---

[2] https://thenew.org/org-people/about-pir/policies/anti-abuse-policy/#:~:text=Technical%20Abuses%20of%20the%20DNS&text=This%20Policy%20prohibits%20the%20following,access%20to%20private%20computer%20systems.

3 https://kenic.or.ke/demystifying-domain-name-system-dns-abuse/ and KENIC Registrar Accreditation Agreement, https://kenic.or.ke/wp-content/uploads/2020/01/KENICS-REGISTRAR-AGREEMENT-1.pdf.

4 People's Republic of China, 2017, Article 3, https://www.cnnic.com.cn/PublicS/fwzxxgzcfg/201710/t20171026_69608.html.

5 Regulation (EC) No 733/2002 of the European Parliament and of the Council (3) and by Commission Regulation (EC) No 874/2004 (4).

6 Resolution CGI.br/RES/2008/008/P.

7 mw ccTLD Domain Registration Policy Version 1.2a of 23 July 2015.

8 Electronic and Postal Communications (Domain Name Management) Regulations, 2020.

9 https://www.article19.org/resources/icann-content-moderation-at-the-infrastructure-level-is-a-dangerous-move/.

The policies of the (.cn) registry have been criticized[10] for being inconsistent with international human rights law, particularly regarding freedom of expression. This criticism stems from the fact that through DNS policies, China is censoring access to information for citizens and residents within Chinese borders. The (.eu) registry has implemented a content removal policy, which has been criticized by this author before for potentially violating the right to freedom of expression and access to information.[11]

The policies of the (.br) registry have been praised for being consistent with international human rights law and promoting transparency and accountability in the management of DNS abuse.[12] The (.mw) registry has developed a domain name dispute resolution policy, which attempts to align with international best practice. We can however note that beyond analysing the policies themselves for these registries, there is lack of clear updated information on implementation of the policies and even in instances where an analysis of the policy notes that the policy is compliant with international best practices, there is still a chance that the implementation might be marred with loopholes to allow for violation of human rights policies.

The (.tz) registry has implemented a policy on domain name suspension and deletion, which has been criticized for lacking transparency and potentially violating the right to due process.[13]

Overall, registry policies have a significant impact on DNS abuse, intermediary liability, and censorship. It is important for registries to ensure that their policies are consistent with international human rights law and best practice to ensure the protection of human rights while mitigating DNS abuse.

Conclusion and Recommendations

In this article, we have examined the intersection of DNS abuse, intermediary liability, and human rights, with a particular focus on the right to freedom of expression. We have analyzed the institutional and legal framework surrounding DNS abuse mitigation measures, as well as the policies of select registries, against international human rights law and best practice.

Our analysis has highlighted that DNS abuse poses a significant threat to human rights, particularly the right to freedom of expression. While DNS abuse mitigation measures are necessary, it is crucial to ensure that they are consistent with international human rights law and do not result in censorship or violate other human rights.

We have identified key stakeholders within the ICANN and Internet Governance ecosystem who are involved in the implementation of DNS abuse mitigation measures and accruing liability. These stakeholders have a responsibility to ensure that DNS abuse mitigation measures are consistent with human rights and to promote transparency and accountability in their implementation.

Our analysis of select registry policies has highlighted the need for registries to ensure that their policies are consistent with international human rights law and best practice. While some registries have

---

10 https://link.springer.com/chapter/10.1007/978-3-031-28486-1_19 and https://digitalmedusa.org/wp-content/uploads/2021/08/Requiem-SSRN.pdf and https://www.tandfonline.com/doi/pdf/10.1080/23738871.2020.1805482.

11 https://www.article19.org/resources/europe-content-moderation-at-infrastructure-level-must-respect-human-rights/.

12 https://isoc.org.br/files/Study_on_the_Marco_Civil.pdf and https://link.springer.com/chapter/10.1007/978-3-030-63501-5_8.

13 https://www.article19.org/resources/icann-content-moderation-at-the-infrastructure-level-is-a-dangerous-move/.

implemented policies that align with human rights, others have been criticized for being inconsistent with human rights and potentially violating the right to freedom of expression.

Based on our analysis, we recommend the following:

1. Incorporate human rights considerations into DNS abuse mitigation measures: DNS abuse mitigation measures should be designed and implemented with a focus on protecting human rights, particularly the right to freedom of expression. This requires collaboration between stakeholders within the ICANN and Internet Governance ecosystem to ensure that DNS abuse mitigation measures are consistent with human rights.
2. Promote transparency and accountability in DNS abuse mitigation measures: It is essential to ensure that DNS abuse mitigation measures are transparent and accountable. This requires the development of clear policies and procedures, as well as regular monitoring and reporting on the implementation of these measures.
3. Conduct regular assessments of registry policies: Registries should conduct regular assessments of their policies to ensure that they are consistent with international human rights law and best practice. Where policies are found to be inconsistent with human rights, registries should take steps to revise and improve their policies.
4. Integrate human rights considerations into business decision-making: The business community has a significant role to play in promoting human rights in the management and handling of DNS abuse. Businesses should integrate human rights considerations into their decision-making processes and ensure that they are not contributing to human rights abuses through their actions or inactions.

In conclusion, DNS abuse poses a significant threat to human rights, particularly the right to freedom of expression. While DNS abuse mitigation measures are necessary, they must be consistent with international human rights law and not result in censorship or violate other human rights. We recommend the incorporation of human rights considerations into DNS abuse mitigation measures, the promotion of transparency and accountability in the implementation of these measures, regular assessments of registry policies, and the integration of human rights considerations into business decision-making. By taking these steps, we can mitigate DNS abuse while protecting human rights.

Centering Victims is Imperative for Effective Remediation in Platform Governance

Thobekile Matimbe
PARADIGM INITIATIVE, LAGOS, NIGERIA

Introduction

The civic space in some African countries is constantly shrinking due to crackdowns on journalists and civil society actors, causing concerns over their ability to express themselves freely.[1] Online gender-based violence is a serious pandemic for vulnerable groups such as women and children. Hate speech is unleashed on marginalised groups in Africa, and harm is perpetrated using social media platforms.[2] Increasing surveillance and arbitrary deployment of legislation to censor and halt the media results in censoring vulnerable groups from online platforms. Platform governance exists in this aggressive context where privacy and freedom of expression are constantly at risk.

A victim-centred approach is key in interventions that promote human rights, whether led by State or non-State actors. Drawing from the United Nations (UN), this approach puts the needs of victims and their safety first, includes a continuous and holistic approach to the delivery of services, and creates an enabling environment for the victims to be heard and supported.[3] Within the context of gross violations of human rights, the UN states that 'victims should be treated with humanity and respect for their dignity and human rights, and appropriate measures should be taken to ensure their safety, physical and psychological well-being and privacy, as well as those of their families.'[4] While quasi-judicial bodies do not have the same binding force as the courts, they can give important guidance in remedial action. In the context of platform governance, platform councils, such as the Oversight Board, may act as quasi-judicial bodies, giving persuasive guidance to the platform while lacking binding authority. The Oversight Board focuses on supporting free expression and other human rights, independently reviewing and making difficult decisions or giving advisory opinions about what content to leave up and take down.[5]

Social media platforms (SMPs) have policies and community standards that define which content is not acceptable on their platforms, and these purport to not only promote business expediency but safeguard the rights of users. Internet intermediaries are non-State actors and not parties to human rights treaties that bind States. Nevertheless, they do have considerable impact on human rights, and there are wide calls across society for affording them according responsibility to promote human rights, examples being transparency of processes addressing content moderation and effective remedial action for victims. Against this backdrop, once SMPs take on a corporate responsibility to protect human rights, platform councils such as the Oversight Board should also apply a victim-centred approach in discharging their function. Looking at the Oversight Board, this approach is lacking as their decisions

---

[1] Paradigm Initiative Londa 2021 Report https://paradigmhq.org/londa/ (accessed 19 February 2023).

[2] Africa renewal https://www.un.org/africarenewal/magazine/all-out-fight-against-hate-speech (accessed 28 March 2023).

[3] United Nations I have the right https://www.un.org/en/victims-rights-first (accessed 20 March 2023).

[4] United Nations https://www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-and-guidelines-right-remedy-and-reparation

[5] Oversight Board.

and advisory opinions are non-pecuniary and have not explored prescribing multistakeholder cooperation in providing psychosocial support for victims of harmful content. Without overstretching the role of the Oversight Board, substantively making their remediation more relevant to victims is necessary in effectively addressing user harm. This approach will also strengthen Meta's ability to consult the relevant stakeholders effectively and strengthen its community standards.

United Nations Guiding Principles

Bearing in mind fundamental rights are non-binding for non-State actors, a human rights lens is still instructive. The United Nations Guiding Principles on Business and Human Rights (UN Guiding Principles) set out the responsibilities of businesses towards human rights, and social media platforms fall within this category. Applicable human rights would then be freedom of expression as is guaranteed in article 9 and article 19(2) of the African Charter on Human and Peoples Rights and the International Covenant on Civil and Political Rights, respectively. However, the right can be limited in terms of a law, in pursuance of a legitimate aim and the means of limitation must be necessary and proportionate in accordance with international human rights standards. The Siracusa Principles and the African Commission on Human and Peoples' Rights Declaration of Principles on Freedom of Expression and Access to Information in Africa (the Declaration) are instructive here. SMPs should be transparent about how they address human rights and the remedial steps they take to address victims of harmful online content.

Remedial action is critical to the role of non-state actors as elaborated in the UN Guiding Principles.[6] The same stipulates the importance of an effective remedy for victims.[7] This approach calls for a victim-centred approach for SMPs and platform councils to take when addressing the human rights impacts arising from the actions of SMPs. Similarly, this approach should guide the adjudication process of platform councils in reviewing the conduct of non-state actors regarding human rights. In addition, the UN Guiding Principles stipulate that in providing remedies for business-related human rights abuses, mandates of existing non-judicial mechanisms can be expanded, citing national human rights institutions as critical.

Access to Effective Remedies

In Zimbabwe, women bear the brunt of online gender-based violence. Many are forced to remove themselves from online platforms. A report on online gender-based violence in Zimbabwe captures informant A's story of how she refrained from Facebook and WhatsApp for a year after her boyfriend shared her nude pictures online.[8] She did not get help from the police and was not aware whether 'Facebook' took down all her nude pictures taken without her consent. Many victims of the non-consensual sharing of intimate images face mental distress to the point where the breach of their privacy and security of person leads them to severe depression. Victims of online gender-based violence may need psychosocial support and effective remedies to address the harm caused. Social Media Platforms

---

[6] UN Guiding Principles on Human Rights and Business, Page 4https://www.ohchr.org/sites/default/files/Documents/Issues/Business/Intro_Guiding_PrinciplesBusinessHR.pdf (accessed on 19 February 2023).

[7] UN Guiding Principles on Business and Human Rights https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (accessed on 19 February 2023).

[8] T Matimbe, Country Report on Zimbabwe Page 63 https://genderlinks.org.za/wp-content/uploads/2022/05/Understanding-Online-GBV-in-Southern-Africa-FINAL.pdf (accessed on 20 February 2023).

can address the prevalence of such cases by taking on advisory opinions that reflect prevalent harm to vulnerable groups and recommend effective remedial action beyond the appeal processes in the select privileged cases. In the case of the Oversight Board, issuing effective advisory opinions on how to meet the needs of victims of online gender-based violence and other harms online can go a long way in effective remedial action for victims. This approach has been deployed by the United Nations Refugee Agency in addressing the needs of victims of sexual misconduct where psychosocial support and risk assessments feature as a remedy for victims.[9]

The UN Guiding Principles aptly provide that situations may arise where a business enterprise requires active engagement in remediation, by itself or in cooperation with other actors.[10] The Due Diligence Project, in a submission to the report by the Office of the High Commissioner for Human Rights (OHCHR) on how to address the gender digital divide from a human rights perspective, highlights concerning online violence that internet intermediaries cannot be held liable for the initial violence, but obligations exist for both state and non-State actors.[11] As such, SMPs must consider the best form of redress for victims in their decisions. In platform governance, the remedies for harmful content that social media councils can order, are, for instance, taking down content for victims of online gender-based violence or reinstatement of content that does not violate community standards or policies.

Consideration of Cooperation Orders

In the Tigray Communication Affairs Bureau case (2022-006-FB-MR), the Oversight Board upheld Meta's decision to remove content threatening violence in the Ethiopian conflict.[12] However, the Oversight Board was concerned about the enforcement of Meta's content policies in times of armed conflict and ordered Meta to look at the feasibility of a sustained internal mechanism that 'provides the expertise, capacity and coordination required to review and respond to content effectively for the duration of a conflict'. While this decision addresses the non-recurrence of harm in conflict areas or attempts to resolve it, reliance on internal mechanisms alone does not suffice. An order for cooperation with other stakeholders in addressing violence online is a more wholesome approach to addressing harmful content.

Multistakeholderism is critical to limit the harms experienced through social media platforms. Consulting extensively with victim support groups and clinical psychologists is part of engaging different stakeholders in addressing victim needs, an approach that can be adopted by platform councils. The United Nations Office on Drugs and Crime (UNODC) highlights 'the right of victims to an adequate response to their needs', which is essentially a victim-centred approach that includes looking at support and assistance as important in remediation.[13] Platform councils must be empowered to give more effective remedies beyond acting as last-resort teams of content moderators where they handle appeals, as in the case of the Oversight Board. When content causes harm, an effective remedy for victims is

---

9 UNHCR A Victim-Centred Approach https://www.unhcr.org/victim-care.html (accessed on 20 March 2023).

10 UN Guiding Principles on Business and Human Rights
https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf (accessed on 19 February 2023).

11 Due Diligence Project Eliminating Online Violence Against Women and Engendering digital equality
https://www.ohchr.org/sites/default/files/Documents/Issues/Women/WRGS/GenderDigital/DueDiligenceProject.pdf (accessed on 20 February 2023).

12 Oversight Board https://www.oversightboard.com/news/592325135885870-oversight-board-upholds-meta-s-decision-in-tigray-communication-affairs-bureau-case-2022-006-fb-mr/ (accessed on 20 February 2023).

13 UNODC Topic three - The right of victims to an adequate response to their needs https://www.unodc.org/e4j/zh/crime-prevention-criminal-justice/module-11/key-issues/3--the-right-of-victims-to-an-adequate-response-to-their-needs.html (accessed on 20 February 2023).

critical, and SMPs should be part of that process of addressing the harm as part of their corporate responsibility.

Social media councils can make decisions even if they are non-binding, but their function as quasi-judicial bodies must permit them to make key recommendations that go beyond community standards in addressing harmful content in line with human rights standards. In *Case 2021-011-FB-UA,* the Oversight Board found the decision by Facebook to remove a post discussing South African society under its Hate Speech Community Standard was appropriate as the post contained a slur deemed degrading, excluding and harmful to the people it targeted, within the South African context. This decision was victim centred as it considered a historical past of a group of people within South Africa and ruled that the slur in issue was against community standards. Beyond removing that content, the SMP decision would benefit from a recommendation of cooperation with the South African Human Rights Commission[14] and organisations promoting transitional justice to raise awareness of the retrogressive nature of hate speech shared on its platforms to foster a culture of human rights.

Leveraging Accountability Mechanisms

Platform councils like the Oversight Board lack the binding force to address the effects of harmful content effectively but can make advisory opinions and preside over appeals. Beyond this, they cannot address systemic online violence, and they need to cooperate with other accountability mechanisms within the different contexts and jurisdictions where harmful content originates. In the context of harm caused by political actors targeting vulnerable groups such as human rights defenders and journalists, working with national human rights institutions with powers to conduct investigations and address the State actors would help reduce the incidence of harmful online content. Cooperating with civil society actors will also ensure that the Oversight Board interacts well with the challenges faced by victims and orders custom-made remedial action that is victim-centric.

Conclusion and recommendations

There is a need for SMPs to adopt a post-content moderation collaborative remediation process that meets the needs of victims of harmful content. This is an innovative approach that SMPs can adopt to ensure they discharge their responsibilities in the human rights protection matrix. Platform councils can lead by prescribing this approach to foster better accountability from SMPs, demonstrating they care beyond financial expediency. Harmful content always comes at a cost, one way or the other. For the private sector, where they retain harmful content in the public domain for longer than necessary, their actions violate a victim-centred approach prolonging the harm caused. Remedial action should go beyond the removal of the harmful content to also address other forms deemed necessary by victims. In cases where platform councils get involved in the appeal of decisions made by SMPs, the victims, as a rule of practice, must be consulted about the nature of remedial action that meets their needs.

There should be consideration of psychosocial support for victims of grave online hate and online violence, for instance, supported by companies and other stakeholders to mitigate the adverse impacts of online content and promote peaceful societies. Companies must be proactive in their corporate social responsibility efforts to build a human rights-respecting environment. While companies are not directly

---

[14] A national human rights institution in South Africa.

liable for the pecuniary cost of online harm, the responsibility to promote human rights is not only for state actors but includes businesses. Platform councils, as non-judicial mechanisms to address content moderation, must prescribe evolving solutions to challenges in accordance with the nature of each case. Their decisions can also foster peace and security by promoting the responsible use of online platforms beyond community standards.

The following recommendations are critical to foster this approach:

– Decisions of platform councils must go beyond takedown of content and recommend cooperation of social media platforms with other accountability mechanisms in addressing the problem of harmful content.
– Platform council advisory opinions must proactively address prevalent cases of harmful content even where appeals are not lodged.
– Platform councils must consider cooperation with other non-judicial and accountability mechanisms to ensure a full appreciation of victims' needs in order to develop their decisions in human rights protection the way judicial bodies will advance their jurisprudence.
– Adequate and effective remedial action must consider promoting peace and security, psychosocial support for victims and other remedial action not limited to the obvious checklist of content moderation.

Platform Democracy, one size does not fit all: the case of GhanaWeb

Emmanuel Vitus
AW FREE FOUNDATION, ACCRA, GHANA

Competing interests among platform owners, users, and governments

In recent years it's increasingly becoming difficult for platforms around the world to identify and deal with harmful discourse. Such content often continues to gain traction online. People responsible for hateful speech online are constantly exploring new methods of bypassing the content moderation tools used by platforms to identify and control hateful speech online. This further complicates the challenges faced by platforms in addressing such content while also respecting freedom of expression. There are no easy answers. This paper looks at a particular example of how these challenges play out on GhanaWeb, Ghana's leading digital news and social media platform.

In Africa, a combination of factors makes content moderation particularly difficult, including colonial legacies, authoritarian governments, and the shrinking of civic space. Once renowned for their feisty media and defense of political liberties, Ghana's journalists and activists are under siege by the police. Online journalists and bloggers increasingly risk arrest and online harassment for their critical reporting.

Internet users in the West African nation can face criminal penalties for online content that is deemed to be false, which is broadly defined under existing law. In June 2021, members of Parliament proposed the Promotion of Proper Human Sexual Rights and Ghanaian Family Values Bill, 2021. If this legislation is passed, individuals who use internet services to produce or share materials advocating or supporting LGBT+ activities would face criminal penalties, including prison sentences of up to ten years. The proposed law was still being considered by Parliament.

In many African countries, due to questionable colonial laws that are now being transplanted into contemporary cybercrime legislation, platforms are now faced with complying with problematic laws that clearly violate free online speech. In Ghana, a new law passed in December 2020 which expands the government's legal authority for surveillance, despite existing data protection policies.

Much grassroots political activism in Africa relies on the use of U.S.-based social media platforms. However, these platforms also play host to state-backed manipulation efforts and can be subject to draconian shutdowns if the political dialogue goes awry for African governments. As a result, these companies get caught up in a tenuous position. To make matters even more difficult, there is an inescapable tension between the platforms' desire to apply global standards, mostly based on algorithms, to content moderation adjudication on the one hand and, on the other, to defer to local contexts when moderating content.

In the midst of this ruckus between platform owners, users and government, GhanaWeb, one of Africa's oldest websites, stands out in its approach to balancing the freedom of speech of its audience and the growing global trends to moderate online content perceived as potentially harmful.

Ghana's leading platform: GhanaWeb

GhanaWeb is Ghana's leading digital news and social media platform. It is an open platform launched in 1999 that operates under the laws of the Netherlands, a legal identity that allows Ghanaians to express themselves freely through opinion articles, multimedia content, and comments. The platform also hosts millions of user-generated posts published through the GhanaWeb Reporter, which is a digital media sharing platform integrated into the GhanaWeb App to give citizen journalists and content creators direct access to the means to publish a wide range of local content.

According to the Alexa website traffic statistics, GhanaWeb is very popular among Ghanaian diaspora in the United States, United Kingdom, Germany, Canada, Italy, South Africa, Netherlands, France, and many other countries.

GhanaWeb's history spans over two decades. It started in 1992 as a print magazine called GhanaHomePage serving the growing Ghanaian community in Amsterdam by collating news stories from Ghanaian newspapers. The GhanaWeb website was launched on January 1, 1999. In 2001, GhanaWeb was the first African website offering a mobile website which was initially supported by feature phones such as the iconic Nokia 3110. Through the Wireless Application Protocol (WAP), GhanaWeb went mobile six years before the first iPhone hit the market.

The portal grew steadily with many people helping to update the website with daily news and information from Ghana. Notable journalists, bloggers and many columnists have gained popularity through their work and opinions on GhanaWeb.

Out of the hundreds of active news and social media platforms in Ghana, GhanaWeb has by far the largest content moderation operation since the 2020 presidential election in the country. The platform's move followed the vote of the Cybersecurity Act, 2020 which provides broad authority for the Cybersecurity Authority (CA) to block or filter online content on receipt of a court order. The law also places penalties on service providers that fail to comply with a blocking or filtering order, which may include one to five years' imprisonment.

In addition, it is the platform that has come under the most scrutiny by the government for its content moderation decision-making practices, both human and automated. One of the reasons for this is that GhanaWeb is the largest news and social media platform in the West African country. It ranks third in national internet engagement after YouTube and Google.com since 2008.

GhanaWeb's content moderation policy

As a result, GhanaWeb's content moderation practices affect a significant amount of user expression. GhanaWeb utilizes both a centralized and hybrid approach to content moderation. However, 90% of the content moderation work is done by the community using a flagging system. The flagging mechanisms allow users to express their concerns by raising a red flag to report offensive or abusive content.

According to the CTO of Ghanaweb, the platform has a strict policy on comment moderation: 'Comments are moderated but not deleted on the platform.' This is in line with the total freedom of expression policy and values of GhanaWeb on comments and feedback from readers.

"*We believe that, at the very least, everyone has the right to be heard. It is not our job to shield the public from unpopular ideas. We provide the platform for as many Ghanaians as are willing to speak out to get their voices heard. Of course, in promoting freedom of expression we are alive to our responsibility to ensure that we do not promote hate speech, criminality, immoral conduct and any forms of expression that will jeopardize national security and cohesion. However, we remain as impartial and*

independent as possible from politics and remain a platform for the voiceless either through contents or comments", said the Editor-in-Chief of the platform.

In his words, GhanaWeb is an independent platform that owes allegiance to no political party, business entity or ethnic or religious grouping. *"Our first loyalty is to our audiences who turn to us every day in search of news and information to help them live their best lives and make good decisions. We don't take sides in any argument or debate. Our job is to ensure that all sides of every argument are laid bare before our audiences to help them make informed choices. We do not give any side more prominence in order to promote any particular agenda."*

Community to Combat Hateful Discourse

In an attempt to ensure that its internal policies can be enforced appropriately GhanaWeb tries to involve the community in a flagging system allowing the community of its over three million loyal visitors to do peer checking of comments and raise a red flag on comments that do not respect the community standard. This includes comments in local languages and icons. It's the first approach to comment moderation. However, once a comment reaches a certain threshold (number of flags), moderators reviewers who receive the same general training on the company's Community Standards and how to enforce them will evaluate the comments and make a decision. *"These guidelines are meant to help the editorial staff and moderators of GhanaWeb live up to our values. They are meant to guide our decisions as to what we keep or moderate and how we moderate them. They are guidelines, not doctrine or dogma. Every situation needs to be appreciated on its own facts and carefully considered on those facts and the discretion of the editorial team. In all situations, moderators are encouraged to have frank and open discussions among themselves with the aim of arriving at the best possible solutions in the interest of all concerned, the audience and the platform,"* said the Editor-in-Chief.

Comments that get blocked can still be visible to those who choose to see them. The system places a caution wall on all comments that are flagged. According to owners of the platforms, commenters don't have to register to comment, they can't contest because they are anonymous. However, those who flag have to register to flag the comments. This makes the content moderation process transparent.

The language bit

The approach to language in Ghana is very important for GhanaWeb, because harmful content continue to gain traction online and those perpetuating it are constantly using new methods of getting around moderation tools, such as using local languages, symbols, memes and emojis to "disguise" the content so that the content moderation algorithm does not identify it as being problematic.

Ghana is a multilingual country in which about eighty languages are spoken. Of these, English, which was inherited from the colonial era, is the official language and lingua franca. Of the languages indigenous to Ghana, Akan is the most widely spoken in the south. Dagbani is most widely spoken in the north.

As part of its hybrid approach to content moderation, the GhanaWeb editorial team engages in several phases of technical and human review in order to identify, assess, and take action against content that potentially violates its Community Standards through using local languages.

Automated content moderation and other methods

In response to growing pressure from stakeholders such as government and the public to take down content that violates community standards quickly, GhanaWeb is now investing heavily in automated tools for content moderation. These include image recognition and matching tools to identify and remove objectionable content such as terror-related content and discriminatory comments directed at particular tribes or communities using language matching tools that seek to recognize and learn from patterns in text related to topics such as propaganda and harm. Depending on the level of complexity and the degree of additional judgment needed, the content may then be relayed to human moderators from the GhanaWeb editorial team.

As time passes, local platforms will continue exploring news methods to address the issue, however, policy-makers should resist the siren call of tighter regulation and illiberal measures. Instead, efforts should be made to mitigate the more damaging effects of social media in ways that take into account local information environments. This could include formulating more targeted measures to improve digital literacy and building trust in key institutions, such as the traditional media, which are often best able to act as arbiters of the truth locally.

The future of GhanaWeb

The case of GhanaWeb shows how local platforms in Africa are struggling to build a transparent and strong moderation system to enhance user expression/experience and also keep the platform safe and democratic. In order to align it's content moderation with the global trends and best practices, GhanaWeb must:

Balance moderation with censorship

Given the scale and reach of GhanaWeb, its content moderation policies need to account for the societal harms that can result from the mass distribution of hate speech and misinformation. The platform owners have a responsibility not to curtail speech too aggressively. Since hate speech and misinformation can be difficult to define, excessively restricting the reach of contentious political speech risks unduly limiting the freedom of expression on which democratic discourse depends.

Publish content guidelines and policies

GhanaWeb should clearly disclose what their content moderation policies are. Ideally, the policies would also be easy for users to understand and include either examples or clarifications of how ambiguous terms will be interpreted. Clear guidelines need to exist about what categories of content or comment will be restricted. Without public and transparent guidelines, content moderation decisions will appear ad hoc and undermine user trust.

Disclose moderation practices and appeals process

GhanaWeb owners should publicly and transparently disclose high-level details about their content-moderation practices, as well as their review process and publish clear guidelines for how to contest a moderation decision: If a comment or content has been banned, users have a right to know how to appeal that decision and whether the review process will involve an automated or manual review.

Disclose Algorithms

GhanaWeb owners should publicly and transparently disclose basic factors about what kinds of data the algorithm considers when flagging or removing a comment or comment on the platform.
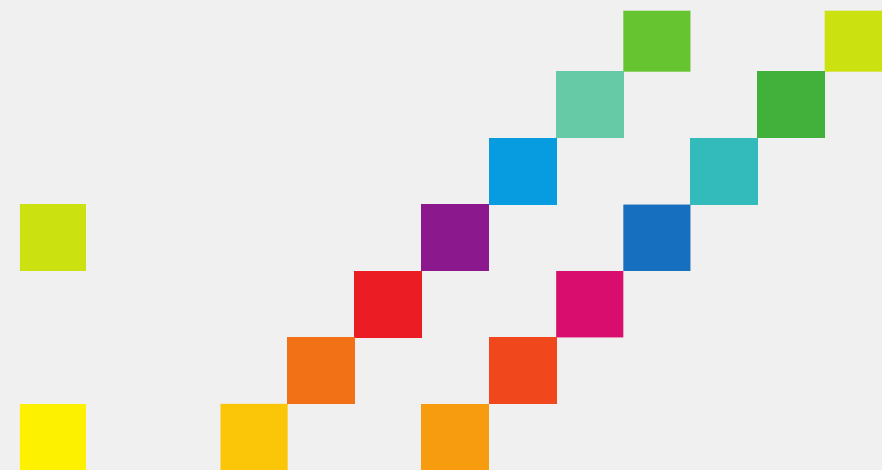
MATTHIAS C. KETTEMANN, HEIDI TWOREK AND JOSEFA FRANCKE (EDS.)

Platform://Democracy

Research Report Americas

Platform://Democracy

Perspectives on Platform Power, Public Values and the Potential of Social Media Councils: Research Report Americas

edited by Matthias C. Kettemann, Heidi Tworek and Josefa Francke

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

Contributors

Name(s)	Affiliation
Heidi Tworek	University of British Columbia, Canada
Luca Belli	Fundação Getulio Vargas Law School, Brasil
Katie Harbarth	Anchor Change, USA
Kate Klonick	St. Johns University Law School, USA
Emma Llansó	Centre for Democracy and Technology, USA
David Morar	Open Technology Institute, USA
Aviv Ovadya	Berkman Klein Center, USA
Peter Routhier	Internet Archive, USA
Fabro Steibel	Institute for Technology and Society, Brasil
Alicia Wanless	Carnegie Endowment for International Peace, USA

Table of Contents

Public Values and Private Orders in Social Media Councils – Perspectives from the Americas

Heidi Tworek
UNIVERSITY OF BRISITH COLUMBIA, VANCOUVER, CANADA

It is a platitude now to emphasize the need to insert civil society voices and public values into social media platforms' decision-making. What is not a platitude is thinking through how to make this happen concretely. The nine papers in this collection from the Americas regional clinic provide concrete and specific insights into how institutions like social media councils might inject public values into the private order of platforms.

Social media councils are not a new idea. The council part of social media councils goes back a century, if not more, depending upon whether we want to count councils as similar to commissions. The social media aspect of councils emerged in what Kate Klonick sees as the second phase of multi-stakeholder involvement in platforms. This occurred from around 2016 with multi-stakeholder intervention rather than the model of multi-stakeholder influence that held sway from 2006 to 2016. Twitter, for instance, announced its Trust & Safety Council in February 2016.

From around 2018, researchers and civil society organizations suggested that councils could not just be based within platforms, but could embed public values and broader concerns around freedom of expression. Pierre François Donquir, then at Article 19, suggested that social media councils could entrench human rights into content moderation processes. At a similar time, I started to write about the potential for social media councils in Canada (what Chris Tenove, Fenwick McKelvey, and I called a "Moderation Standards Council") or as one institution that could work on both sides of the Atlantic to resolve disputes around content moderation in a fair, accountable, independent, transparent, and effective way.

Much of this discussion initially intertwined with debates around Germany's Network Enforcement Law (NetzDG), which came into force in 2018. Some considered whether there were other methods beyond government requirements for social media companies to take down speech flagged by users as potentially illegal within 24 hours. Others highlighted the many speech issues that takedowns could not solve, such as the intransparency of recommender systems.

As more regulation has emerged, including with the Digital Services Act in the European Union, the discussion around social media councils has continued. Highly prominent self-regulated councils, principally the Facebook Oversight Board, also came into being. This sparked new debates about the limits of both self-regulation and global content moderation councils. It also highlighted the ultimately private nature of the Board, which contrasted with the spirit of suggestions for social media councils to incorporate broader public values.

The nine papers of this collection provide nuanced and practical views on how to incorporate those public values into suggestions for social media councils, now that we live in a world of increasing regulation and platform self-regulation. As the papers show, there is much work to be done.

Fabro Steibel's paper points out that policy framing around social media councils could shape their values and the institutions themselves. Values such as human rights and digital constitutionalism have discursive power; how we talk about the values behind social media councils can shape their institutionalization.

If we discuss social media councils through public values, then, we might also remember the contested nature of the idea of a "public" or to use the classic phrase "public sphere." Jürgen Habermas' ideas around the evolution of a "public sphere" in the 18th century has long been criticized for focusing on white, middle-class men. The critiques of Habermas over the last thirty years are one example of how definitions of publics and public values evolve over time. Feminist and class critiques in the early 1990s, a few years after Habermas' work was translated into English, noted that women and working-class milieus formed publics and held their own values too, even if they didn't fit into Habermas' normative framework. More recently, Wendy Willems has pointed out that Habermas' concept of a bourgeois public sphere elided slavery and colonialism, reproducing racialized hierarchies without even mentioning them. The silences in the canonical text showcase the importance of not assuming that any one voice defines "the public" or public values.

Peter Routhier's and Aviv Ovadya's papers take up the challenge of inserting the public in social media councils in two different ways. Ovadya's paper takes a pragmatic approach to inserting democracy into platforms through processes of deliberative democracy such as sortition or temporary advisory bodies. Drawing from a broader literature on democratic processes like citizens' assemblies, Ovadya suggests how platforms themselves could democratize their processes, even before public institutions step in to regulate.

Meanwhile, Routhier's paper suggests that we continue to think more broadly about public interest values as a constitutive part of our online lives. He warns that in the worst case, social media councils could legitimize and entrench current social media platforms. Rather than rely solely on councils, Routhier provides broader ideas of how to embed public interests into the internet, for example by empowering libraries and non-profit sites like Wikipedia.

Although social media councils could offer the chance to bring publics into private orders, Emma Llanso, Katie Harbath, and Alicia Wanless provide important insights into areas where expertise may be necessary. Llanso's paper considers whether and how to incorporate technical expertise into any such councils. Most people, including researchers and policy-makers, may not understand the technical infrastructure of platforms. Llanso offers practical ideas on how to include technical expertise into a council's design and operations. Meanwhile, Katie Harbath's paper sees councils as a space of bridge-building, potentially between platforms and civil society. This could be a crucial role for a council, given that only employees and former employees have access to certain forms of knowledge.

Wanless offers a complementary perspective on the challenges of building councils or council-like bodies in a participatory way. She reflects on the (ongoing) process of creating a CERN for the information environment. Such an international body would foster interdisciplinary research, so vital given the vast number of open questions around platforms' influence on democracies. But it also presents challenges of how to build an entity that is bottom-up, while still having some coherent direction.

The final three papers consider the implementation of any council model. David Morar points out the limitations of any models that do not embed sufficient capacity for enforcement. Using examples from other industries, Morar lays out the pitfalls of weak enforcement, including greenwashing or bluewashing ("functionally toothless labelling schemes" on human rights and labor standards). Morar demonstrates the necessity of enforcement mechanisms, while suggesting that any council will find it hard to enforce its decisions if it does not incorporate platforms in some way.

For Luca Belli, such considerations help to explain why regulation will be required to insert public values into platforms. To solve problems, Belli argues, legislators would have to design social media councils to be "meaningfully accountable" instead of self-regulatory. Finally, though with very different

caveats, Kate Klonick warns that more voluntary social media councils may swiftly fall victim to cost-cutting during economic downturns. Her paper thus suggests mandates for any public-private versions of social media councils to avoid sudden dissolutions of voluntary bodies.

Overall, the nine papers from the Americas regional clinic tackle the practicalities of the oft-proclaimed idea that publics deserve to be involved in platforms' decision-making. Some papers offer pragmatic solutions for the present, while others point to future modes of institution-building. Still others remind us of pitfalls to avoid. They do not pretend that there is a panacea to the privatized order of the present. And they do not come to a consensus on one route ahead. But they show that there are many more paths forward than the pessimists might postulate.

Platform Councils: Solving or Creating Regulatory Vulnerabilities? A Brazilian Perspective

Luca Belli
FUNDAÇÃO GETULIO VARGAS LAW SCHOOL, RIO DE JANEIRO, BRASIL

An old Soviet adage recommends that "If you do not want to solve a problem, then create a commission." The hypothesis of this brief paper is that platform councils may not necessarily solve digital platforms' deficit of democratic values accountability, and they might contribute to creating further regulatory vulnerabilities, unless legislators design them to be meaningfully accountable.

To this end, this essay provides a brief and non-exhaustive introduction to the types of platform councils developed so far. Subsequently, it explores the existing Brazilian platform regulation framework, and utilises the three main democratic values included in the title of the Brazilian Bill for Platform Regulation i.e., "freedom, responsibility and transparency", to analyse how far platform councils might be useful to foster democratic values. Lastly, the conclusion highlights some relevant caveats, regarding the effectiveness – and, ultimately, advisability – of relying on self-regulation versus public regulation to establish platform councils.

Platform councils and their diverse rationales

As highlighted by the Platforms and Democracy position papers, several types of digital platforms, besides social media, have been experimenting with platform councils over the past five years. The most renowned example is Facebook's Oversight Board, which has established such a body to – supposedly – regulate the content moderation practice of the platform in a more participatory and accountable manner.

That is by no means the only existing example. Platform scholars are aware of gaming platforms experimenting with player councils, such as EVE Online's Council of Stellar Management, a group consisting of ten EVE Online players democratically elected by community of game players, to advise and assist the continuous development of the platform, provide analysis, share suggestions, and give feedback.

A less-known instance is the one created by the Chinese data protection law, the Protection of Personal Information Law (PIPL), whose article 58 establishes an obligation for large platform providers – defined as those having "a huge number of users and complex business models" –to establish an "independent supervision board" composed by external members, responsible for monitoring the correct implementation of the law.

This normative provision represents a remarkably original feature among data protection frameworks that, together with the large platforms' obligation to periodically publish reports on their data processing activities, also foreseen by PIPL article 58 aims at increasing accountability through what is described as "society's supervision."

The increase of platform accountability, in truly meaningful ways, ideally embedding democratic values and oversight in these private entities, has indeed been a recurrent preoccupation of scholars and policymakers alike for the past decade. Such preoccupation has been repeatedly expressed, for instance, by Brazilian policymakers that, since 2020, have been discussing policy efforts aimed at regulating digital platforms, although without reaching a consensus on how to do so, at the time of this writing.

Contextualising Brazilian evolutions

An important element of context regarding Brazilian digital platforms regulation is that social media platforms are already regulated by the Brazilian Civil Rights Framework for the Internet, Law n. 12,965/2014, a.k.a. "Marco Civil da Internet" (MCI), which the Brazilian Congress wishes to supplement with Draft Bill n. 2,630/2020, on Freedom, Responsibility and Transparency on the Internet, a.k.a. the "Fake News Bill."

The MCI is Brazil's primary law regulating the Internet and the first and only general law for Internet governance and Internet rights adopted in Latin America. MCI article 19 establishes a general regime of a judicial notice-and-takedown system where "application providers," i.e. platforms, are deemed liable for user-generated content only if failing to comply with court orders for the removal of specified content within 24 hours, granted they have the technical capacity to do so.

Hence, the Brazilian legislator has framed an approach to regulate digital platforms, although this initial approach has shown limits. The constitutionality of MCI article 19 is currently challenged at the Brazilian Supreme Court level. Plus both the Brazilian Legislative and Executive powers are actively pursuing efforts to supplement the MCI provision with specific normative provisions aimed at improving "freedom, responsibility and transparency".

The Brazilian policymaking initiatives do not mention councils, but for our discussion the Brazilian experience is particularly relevant, as it allows us to understand the democratic values that the Brazilian nation, through its democratically elected Congress, considers as the most relevant and in need of protection, when it comes to regulating platforms.

Democratic values, which could be supposedly baked into platform governance architectures through platform councils, are a very large and heterogeneous set of values. The Brazilian Legislator deems "freedom, responsibility and transparency" as so relevant for platform regulation – and clearly missing from platforms private orderings – to include them in the very name of the Bill that aimed at regulating them. This inclusion leads us to assume that such values are the most relevant ones from the Brazilian perspective.

It is not clear, however, how and to what extent the establishment of platform councils could strengthen the aforementioned values. To provide an initial answer to such query, the following section will briefly analyse how platform councils could be used and designed to positively contribute to improving Brazilian democratic values.

Platform councils: Regulatory trick or treat?

For the sake of clarity and conciseness, this essay will only examine three specific elements of the entire spectrum of democratic values – i.e., freedom, responsibility, and transparency – based on the importance that they seem to play in the Brazilian context. This section provides some observations and recommendations on how each of these democratic dimensions could fit into platform governance through the action of a platform council.

Freedom

This first item evokes the full spectrum of fundamental freedoms granted to each and every individual at the domestic level by national constitutions and at the international level by binding international law frameworks.

From a pragmatic standpoint, it seems highly unlikely that a sound protection of fundamental freedoms can be granted by solely relying on global international law frameworks, such as the ICCPR or ICSECR, as this type of international law framework inevitably relies on national systems to specify and implement normative provisions and, critically, their exceptions. Importantly, the subject of international law obligations are states (i.e., public bodies which must guarantee the full enjoyment of rights to physical persons). Thus, international jurisprudence, while offering important guidance, might be of limited use when trying to establish how corporations (i.e., legal persons) must respect individual rights.

Regional fundamental rights frameworks exist, and in some regions, might be more active than others, even regulating the behaviours of corporate actors, but they are typically renowned for their lethargic processes. In this perspective, the establishment of regional platform councils might be an interesting option to translate the existing regional approach from fundamental rights to guidelines for responsible platform behaviour, as I will discuss in the next point.

Lastly, national constitutional law and domestic jurisprudence are usually well suited to address the specificities of local culture, particularly as regards local juridical sensitivities and traditions. Therefore, if any national platform council must be established, I suggest they should be created at a national level, as convincingly argued also by other authors. Such national bodies could coordinate at the regional level if needed, considering that most international disputes may typically occur at the regional level.

Responsibility

As suggested in the previous paragraph, one of the core elements of a responsible behaviour from private entities in general and digital platforms in particular is the respect of fundamental rights, but also the provision of effective remedies, strongly recommended by the UN Principles for Business and Human Rights.

In this respect, platform councils might serve as a useful additional mechanism for users to seek redress when any of their rights is unduly violated. Considering the existence and notable advancement in terms of sophistication of Online Dispute Resolution (ODR) mechanisms, it seems that national platform councils could be designed to be appellate bodies of the existing ODR mechanism, so that ideally users are granted the most effective and just resolution of their potential controversies.

The ancillary benefit of establishing a system of a national platform council acting as an appeal body would be the decongestion of national juridical systems as regard platform disputes. This could simultaneously improve the full enjoyment of platform users' rights, increase the (corporate social) responsibility of platforms, and positively contribute to alleviating notoriously overburdened national judicial systems.

Transparency

This point deserves special attention as transparency is frequently touted as a sort of silver bullet, able to solve – or at least contribute to solving – a wide range of issues. But de facto transparency may not be as effective as we might think, being usually very poorly defined and even more poorly implemented.

A telling example in this regard are platform terms of service, which should supposedly be a tool of transparency – and indeed are considered so by data protection frameworks around the world. Yet, these end up as instruments cleverly engineered to mislead and confuse the user with lengthy and highly technical terms, without providing any meaningful information that could increase accountability for the provider and oversight for the user – and society.

In this respect also the establishment of platform councils may be helpful, should such organs be mandated to publish regularly reports on specific issues – e.g. content moderation, product defects, game features, etc – that enable an increased understanding, monitoring and, ultimately accountability of these private actors. Importantly, to achieve a situation of <u>meaningful transparency</u>, such reporting should also adopt shared – ideally standardised – formatting requirements so that data can be more easily compared and studied by regulators, researchers and users, and even be machine readable.

Conclusion

While a good dose of scepticism regarding the potential benefits of platform councils must be of order, this brief analysis has demonstrated that platform councils might prove to be useful, largely depending on how they are designed. A system composed of national and regional platform councils might prove to be an interesting choice as they could usefully assist in guaranteeing users' rights and provision of effective remedies, as well as improving platform transparency and, consequently, accountability.

However, a very strong caveat is that such system – and whatever other platform council system – would have a cost. Such a cost might be easily borne by large platforms, as explained by the Chinese legislator's choice to only target large platform with its obligation to establish "independent supervision boards." Small players and, particularly, new entrants in the platform business would be very unlikely to have the resources to establish such a complex and costly system and, therefore, stakeholders' expectations as to the impact of platform councils should be lowered to a minimum, adopting the most pragmatic approach possible.

Lastly, stakeholders should be aware that should the design and implementation of platform council be delegated to self-regulation, with no specific indications from legislators on how they should be established, their roles, responsibilities, and accountability, platforms might use such bodies as a strategy <u>to avoid or postpone important regulatory action</u>.

Indeed, as long as national law does not mandate the establishment and regulation of such bodies, any large platform executive would have a fiduciary obligation towards shareholders – which exist in every country for every publicly traded company, such as most large platforms – to prioritise the maximisation of their profits rather than the full enjoyment of user rights. In this perspective, it seems reasonable to posit that national platform councils regulated by domestic law could play a useful role, improving platform governance, while it seems at best naïve to argue that global platform councils, purely based on self-regulation will be able to offer any meaningful solution.

How Platform Councils Can Bridge Civil Society and Tech Companies

Katie Harbarth
ANCHOR CHANGE, WASHINGTON D.C., USA

Meta's Oversight Board is a first-of-its-kind experiment where a platform council can overturn the company's decisions on content. However, its powers are limited to individual pieces of content, though the Board can make policy recommendations to Meta and serves an important oversight role to hold Meta accountable for its promises.

While the Oversight Board is first in many ways, in others, it is not. For over a decade, platforms have pulled together groups of individuals with various backgrounds to provide guidance and expert advice on content policies, product development, human rights standards, regional and cultural differences, and many more. While these groups can counsel companies on their work, they have no power to force them to do anything. This lack of checks and balances on the companies and their decisions lies at the crux of many of the arguments debated today about holding tech companies accountable.

Moreover, global civil society groups have documented ways platforms can harm society for many years. They regularly lament that the platforms do not listen to them, nor do they even know how to contact someone who works at each company.

Meanwhile, the platforms struggle because if they are small to medium size, they likely don't have even one person dedicated to being a liaison with the civil society community. Even large platforms with multiple employees will struggle to manage requests from the thousands of groups around the world who want to talk to them.

This is where platform councils can play an important bridging role between the expertise of these two communities. These councils can act as a connector, help to translate civil society concerns into things that tech companies understand how to adopt, help civil society understand what's possible technically, help mediate competing priorities, create frameworks for engagement and provide valuable oversight on both communities.

This paper will explore all the roles platform councils could play and what they would need to be effective.

Unique Role of Platform Councils

Building a platform that protects freedom of expression but protects people from harm is a simple thing to say but an incredibly complicated balance to put into practice.

It goes beyond debating what a platform's policies should be on the types of content and behavior they do or do not allow. It also encompasses the following:

- What should the penalties be for violating the rules?
- Should there ever be exceptions, such as for newsworthiness?
- What the local law requires.

- How companies should respond to pressure from the government to remove content – especially when employees' safety might be at risk if they do not comply.
- What to do about borderline content that doesn't violate but is still problematic?
- The platform's ability to enforce said policies – especially at scale.
- The design of the product itself – including how machine learning algorithms are built.
- Transparency into the decision-making processes and an ability for researchers to study how the platforms work and their impact on users.
- And many more aspects.

Debates have ensued for years about how various platforms should answer these questions. Platform councils can potentially play a role in settling these debates by playing the role of:

- **Bridge**. Often, civil society organizations do not know how to engage tech companies – nor have the resources to talk to all of them, and vice versa many tech companies don't know how to reach out to civil society organizations nor have the resources to handle incoming from everyone who might want to engage them. Platform councils can help bridge this gap by organizing civil society concerns in a way that those at the platforms can easily digest. They can also help the tech platforms explain more about how they operate to the civil society organizations.
- **Oversight**. It's not enough for any organization to just promise that they will do something. They need to be held accountable on how they follow through. Platform councils can stay on top of tech companies and civil society groups to understand if their actions match their promises. Regular transparency reports can also help people understand how different companies compare in their efforts and if they are making regular progress.
- **Frameworks**. One of the biggest challenges for both civil society and technology companies is that they don't know how to speak to each other in a way that each will understand. Civil society often doesn't understand how the platforms work, and tech employees don't necessarily understand the human rights world. Platform councils can help to develop frameworks to facilitate conversations and collaboration. An example of how this works could be how platforms prioritize which elections they will work on. A platform council could work with tech companies to understand the various considerations and resources they have to work with and how they prioritize this work today. A platform council could make recommendations on various data sources the companies should use – such as Variety of Democracies – as well as other considerations based on input from civil society.
- **Reconciling priorities**. While better collaboration is needed, that doesn't necessarily mean that everyone will agree on the prioritization in which certain issues are worked on at the companies. Not everything can be done at the same time and reasonable people will disagree on where to draw the lines. Platform councils can play a role in reconciling those differences to help provide guidance on which things are more pressing than others.

To be effective in playing this role described above, the structure and scope of the platform councils must also be considered. Some crucial aspects would be needed:

- **Across platforms**. These problems are not siloed to any platform. To be effective and to ensure consistency where needed a platform council would need oversight and cooperation across various tech platforms and companies.
- **Product and policy oversight**. One major flaw in the Meta Oversight Board structure is that they only have the ability to tell the company to leave up or take down a particular piece of

content. They can make policy recommendations, but the company does not have to follow them. Nor can the Board make any recommendations on the product itself, such as how it is designed or works. To be effective platform councils would need to be able to participate in these areas too.

– **Diversity of backgrounds and skills.** The final important thing would be the make-up of the council itself. The people must be diverse not just in race and gender but also in geography, ideology, and skill sets. You need human rights experts, former platform employees, former government officials, journalists and others to adequately balance decision making.

– **Financing.** Both the platform council and civil society group would need to be compensated for this work. Ideally, funding would come not just from the companies themselves, but also from other philanthropic money. Government funding from USAID or similar agencies could also be considered. These funds would be put into a trust not controlled by the companies that could then fund the work of the platform council but also be used as stipends for the civil society groups who participate.

While platform councils have been around for a while as advisory groups, only recently, with the Meta Oversight Board do they have any power to overturn the decision of the platforms. Even that is very limited so councils do need to be reimagined and given real powers to be effective. By doing this a platform council can provide many meaningful roles, including being a bridge between the expertise that the tech companies have and civil society. By serving in this role they can fill a gap that continues to persist to this day.

The Evolution of Social Media Councils

Kate Klonick
ST JOHNS UNIVERSITY LAW SCHOOL, NEW YORK, USA

Introduction

Jack Balkin's Free Speech Triangle, provides a useful heuristic for understanding the old and new powers in tension to control free expression. Before the internet, the old model of free expression in a democratic society, he argues, was not a triangle at all, but a dyadic model between the State, which threatened to censor using its police power, and its citizens, who used voting, voice, exit, and protest as a means of pushing back against that power. The internet, Balkin argues, ushered in a new corner – and hence the triangle – radically altering the power structure between State and citizens. In this "new school" world, online private speech platforms diminished the power of the state to censor, allowing users to route around such controls via these new communications technologies. But this shift of routing around law was not only in citizens' favor; it also enabled States to circumvent limits on their ability to surveil citizens by using private platforms.

The Free Speech Triangle is useful because it also provides a framework for mapping the current limitations of each corner's power, and how the dangers of certain remedies can exacerbate power inequities particularly for the group we want to empower the most in a democracy: users/citizens.

For instance, users/citizens have little legal recourse against private online platforms, though they do ostensibly have the power of protest, boycott, and reputational shaming. An easy solution might seem to be to empower the law against online platforms on behalf of users/citizens. The irony is that doing so often requires state regulation of speech – that is, using the state to police the online platforms. But state control of speech brings us straight back to the dyadic model, pre-triangle. Especially when coupled with the private surveillance power that platforms enable, state control of platforms is an incredibly dangerous proposition. Indeed, to avoid regulation which might damage their business model, platforms might be more willing to cooperate with governments, despite potential harm to users. Though calling for regulation of a powerful industry might seem a sensible and obvious move, in fact, driving these two nodes of power together might be the most dangerous outcome for users/citizens.

What the triangle reveals is that the problem we're trying to solve for – empowering users/citizens voices and their say in the private structures that control and govern those voices – is maybe not best answered by traditional solutions of regulation through governments. Instead, a new kind of solution might be necessary to strengthen users/citizens' existing powers against private platform governance. This could take the forms of markets, norm enforcement, or, as this essay will argue, the creation of institutions to stand in for the direct representative capacity of individual users/citizens. One such mechanism to do this is social media councils, multi-stakeholder constituencies that themselves reflect the new pluralistic environment within which speech rights now exist.

This paper will reflect briefly on the history of social media councils and their varied successes in fulfilling this role as a point of leverage for users/citizens on speech platforms. It will conclude with a few key lessons from these experiments.

History of Formal and Informal Speech Platform Relationships with Social Media Councils

Multi-stakeholder influence has existed at platforms since the very beginning of social media. For the sake of simplicity, I'll roughly group the nature of this influence into two eras: the 2006-2016 era, which I'll call early "Early Web 2.0" and then 2016 to present, which I'll call "Current Web 2.0."

Early Web 2.0 Multi-stakeholder Influence 2006-2016

For the large and now dominant user-generated content and social media platforms – Facebook, YouTube, Twitter – 2006 is a crucial year. In 2006, Facebook first became publicly available to a global audience and YouTube was created. While MySpace, LiveJournal, Orkut, Friendster, and many other platforms already existed, YouTube and Facebook are the most relevant examples to draw from because:

- They are still dominant user-generated content platforms making them highly relevant to both government and private entities;
- They are still dominant social media platforms making them highly visible in media and society;
- They operate globally;
- For better or worse, their content moderation practices have become the most followed models for U.S. social media companies.

As early trust and safety employees at Facebook and YouTube encountered issues around content moderation for non-intellectual property content, they developed standards based on American norms around freedom of expression. These later developed into more formal rules (and exceptions to rules) formed recursively in a common-law like pattern in response to new fact patterns of content moderation that presented themselves over time.

Though the early internet is often characterized in tech utopian terms by many early scholars and technologists for the freedom of speech and access to information that it provided, research and reflection has shown that the "Wild Wild West" of the early internet also enabled a massive amount of harmful behavior. Unfortunately such behavior – hate speech, harassment, stalking, doxxing, racism, defamation, misogyny, antisemitism – was underreported and underacknowledged by many high-profile academics and policy-makers at the time, in part, unsurprisingly, because such bad acts targeted historically vulnerable populations.

But two groups of individuals had front row seats to these ugly sides of the internet: the individuals at platforms working to remove such content or prevent its posting, and civil society groups that already specialized in helping individuals exposed to such harms in their offline capacity.

It took time, but eventually these groups found each other and began to become informed by each other. In particular, well-funded and respected organizations like the Anti-Defamation League, Electronic Frontier Foundation, and ACLU had early impact in bringing together individuals working on the internal content moderation policy at platforms and helping them collaborate to form best practices. Those in charge of content moderation and trust and safety at platforms were eager for external advice on how to set community standards for their sites and manage trust and safety flows for a number of practical reasons:

- Generally speaking these employees' roles were relatively low profile in the company, and at the time operated independently. Their reasons for joining tech platforms were social planner motivated and not tied to or in answer to revenue generation directly.

- These employees lacked both practical, legal, and philosophical expertise in these areas of harmful speech, or how to make tradeoffs between safety and principles of freedom of expression and democracy. They welcomed input from representatives of people affected by their policies and rules.
- As the influence of content moderation grew along with the size and scale of the platforms, recognizing and accepting input from expert groups seemed not only a practically helpful but more legitimate means of setting global speech policy.
- At scale it was easier and more efficient to speak to civil society, academic, and government stakeholders who served as gate-keepers, than to individuals that were affected by such policies.

Though early collaborations between such groups and platforms were at first informal, platforms formalized these relationships both by creating group councils and by formally empowering individual stakeholders.

Group Councils

Formalization of multi-stakeholder groups came in the form most recognizable today at Social Media Councils. The role and public face of such councils varied. Some, like the Twitter Trust & Safety Council, were private groups of expert individuals convened by the platforms to be on call to answer questions on particularly problematic pieces of content or to review forthcoming changes in community standards or internal speech policies. Others, like the Facebook (now Meta) Oversight Board, was more public-facing and included both a private and platforms process to request review and input on speech decisions and policies. The work demanded by these councils was often both reactive (providing input on particular pieces of content that had already been flagged as problematic which needed expertise in making a decision) and proactive (providing input on proposals to change the speech rules of the platforms in the future).

Individual Trusted Flagger and Partners Programs

Trusted flagger or trusted partner programs were also developed at YouTube and Facebook to do the same type of work as Group Councils but not in a committee format. Rather than group collaboration, such programs were lists of individuals who could be queried – or proactively inform the platforms – in a one-on-one capacity of issues arising on the site. Identification of such trusted flaggers and partners was both meritorious and legacy driven. Trusted flaggers ranged from

- Users who had established themselves as reliable signallers of problematic content or content erroneously flagged for removal by platforms over time;
- Former trust and safety employees of the platform;
- Government officials;
- Expert individuals from academia and civil society whose groups were flagged as "trusted" or who had existing relationships with platforms in other capacities;
- Members of the formal Group Councils in their individual capacity.

There are benefits and drawbacks to these more informal individual relationships. On the one hand, they are easy, cheap, fast, and reliable signal mechanisms for platforms. But such informal networks lack transparency, deliberation, and are prone to cronyism and nepotism.

Though both group councils and trusted partner programs still exist, the growing public attention and awareness around content moderation have led to both popular and government pressure for more transparency on such programs and increasing reliance on more formal councils.

There has also been more active engagement and development of civil society efforts to engage platforms through both these formal and informal mechanisms. This was due in part to the media attention and journalism around content moderation, but also a number of high-profile academic works that highlighted the realities of these pluralistic systems at work, making them more of a target for impact and intervention.

Current Web 2.0 Multistakeholder Intervention

A turning point in both journalism and academic work in this field occurred around 2016 when a number of events happened simultaneously to raise awareness:

- Publication of a number of foundational academic research texts describing the field of content moderation and the pluralistic forces that impacted it;
- High profile examples of both "erroneous" removal of content by platforms and platforms failure to remove other kinds of problematic content;
- The rise in this public, government, academic, and media awareness during a key U.S. presidential race, which had a controversial outcome that many blamed on social media.

The sudden scrutiny of content moderation policies and practices at platforms threatened to hurt the brand reputations of companies and accordingly, their stock valuations. It also threatened to disrupt the user engagement model that funded the advertising revenue on which such platforms were reliant. And of course, content moderation controversies created threats of litigation (however unsuccessful they might be due to Section 230 immunity) and more worryingly for platforms: regulations.

In order to stem some of the reputational damage and user dissatisfaction – as well as pre-empt potential regulation – many social media platforms decided to invest in more self-regulation and to do so more publicly. That took the form of more transparency around speech rules, increased cooperation with outside researchers, academics, and journalists, revamping of content policies and process – and most relevantly for this paper, increased investment in governance structures and institutions, in particular formal social media councils. Facebook's creation and investment in the Oversight Board to review appeals from users on speech removal and non-removal, is perhaps the most high profile example.

Current Moment and the Lessons from the Past

Despite the call for and prevalence of formal social media councils in the last seven years, it is unclear what the long term future holds. A number of factors have changed the environment which – at least temporarily – made platforms voluntarily engage in social media councils. They include:

- A decline in market valuation of tech companies, resulting in widespread layoffs, hiring freezes and other cost-cutting measures, which include the dissolution or elimination of many social planner governance and research cooperation programs;
- The sense of relative lack of impact that platforms governance investments – or "governance washing" – did to repair harm to brand from content moderation scandals and election controversy;
- The failure of self-regulatory efforts like social media councils to stave off regulation like the European Union's Digital Services Act and Digital Markets Act.

For those optimistic about the potential for new public-private institutional development like social media councils to protect users/citizens rights and represent their interests, these developments seem particularly disheartening. But the history of how they developed to this point and the environmental changes that led to their relative rise and fall can illuminate some valuable new tools and lessons to either preserve existing councils, encourage wider adoption, or empower similar pluralistic mechanisms on behalf of users/citizens. These include:

– State regulation of social media and speech platforms is particularly fraught where such platforms enable users/citizens rights around freedom of expression, but also perpetuate certain harms. Empowering the state to regulate platforms must be careful to preserve and protect the rights of users/citizens, not consolidate or strengthen, the power of government and the platforms against users/citizens.

– To this end, the best mechanisms for state power on behalf of users/citizens rights are those that promote process, users/citizens participation in platform governance, and transparency – rather than any substantive regulations around speech.

– Multi-stakeholder input to platforms on their speech policies and decisions developed organically because the relationship was mutually beneficial to both the interests of civil society groups – particularly in the United States and Global North – and the interests of platforms.

– Despite this mutual benefit, platforms have very little economic incentive to formalize or publicize such relationships with multi-stakeholders in any capacity, but this is particularly true for social media councils.

– If investment in social planner initiatives like social media councils has little to no economic benefit for the company and are only cost centers, they will quickly eliminate "self-regulatory" programs in economic downturns.

– Given these realities, mandated creation of social media councils for platforms is likely the only and best route to their continued existence and the best hope for continued development of new public-private institution building that empowers users/citizens.

Technical Difficulties: Incorporating independent technical expertise into platform council decision-making

Emma Llansó
CENTRE FOR DEMOCRACY AND TECHNOLOGY, WASHINGTON D.C., USA

Why was a cartoon critiquing police violence in Colombia suddenly removed from Facebook 16 months after it was first posted? What sorts of safeguards can Meta put in place to ensure that clearly labeled breast-cancer awareness images are not removed as violations of its anti-nudity policy? How should a service use automated filtering tools to swiftly enforce rules in a crisis situation, and what kinds of after-the-fact remedies should users have access to? These are just a few of the questions raised by actual cases that have been taken up by the Meta Oversight Board, established in 2020. They help to illustrate the significant technical component of many of the questions that the Board, or any variation on a platform council, will address about how an online service provider's systems are affecting and reflecting public values.

Reaching useful answers to these questions—investigating the provider's systems, evaluating their impacts on human rights, developing effective and implementable recommendations—will require the evaluators on such a council to have access to a wide range of expertise, including everything from interpretations of international human rights law to the lived experience of the people most directly affected by the online service provider's decision making.

For the purposes of this essay, I focus on the challenge of incorporating technical expertise in the design and operation of a service into the deliberations of a platform council. I identify a tension between two key features of this technical expertise: that it should be pertinent or specific to a given online service, and that it should be provided independent of that service. I also discuss the challenges of confidentiality when working with technical information pertaining to a specific service, and the role of transparency in ensuring both the accuracy and the legitimacy of council decisions. I then explore several options for making technical expertise available to the members of a platform council.

Finding Technical Expertise

Running an online service that provides a platform for user-generated content involves multiple technical systems: systems for carrying out the basic functions of the service, including hosting content, authenticating logins, and enabling accounts to interact with one another; systems for sorting, organizing, finding, and recommending content; and systems for detecting, evaluating, and enforcing moderation decisions against content that violates the service's policies, to name a few. Each of these systems, operating individually and in conjunction with one another, has the potential to affect the human rights of the service's users and may come under the scrutiny of a platform council.

Thus, a platform council will need access to many kinds of technical expertise. It requires different technical training to adequately evaluate, for example, whether a service's approach to moderating hate speech disproportionately affects users from a certain background, or whether the service's content recommendation algorithm routinely suppresses certain topics or perspectives. But technical expertise in the operation of online services is not seen as a primary, or even necessary, qualification for a member of a platform council. Among the Meta Oversight Board members, for example, the most common areas

of expertise include human rights, law, freedom of expression, and policy; Board members generally do not have a background in computer science or in working in trust & safety at a social media service.

Expertise in human rights law is certainly vital to evaluating online services' impacts on users' rights, but it illuminates only part of the picture. Moreover, while it can be helpful to have a general knowledge of how technical features such as machine learning classifiers or recommender systems work, most of the questions that will go before a platform council will be specific to the details of how a particular service operates. There is no single off-the-shelf technical approach to content moderation or recommendation that is employed across platforms. While some of the tools and techniques may be used in common across various services, each service modifies and implements them in unique ways, typically by incorporating them into a bespoke system. The questions of how Meta's systems affect its users' rights are specific to Meta's systems, just as they will be for any individual platform.

So not only do platform councils need access to subject-matter specific technical expertise, but it likely needs to be platform-specific as well. Much of the most relevant expertise will lie within the online services themselves, among the staff who have worked to build and run the very systems being evaluated. The field of online trust & safety is relatively young, and unlike in other technical industries, there do not yet exist massive consultancies of experts with decades of experience in the field who are available to weigh in as independent advisors. (There are several recent initiatives that seek to bring trust and safety professionals together to foster the development of the field (e.g. the Trust & Safety Professional Association) and to contribute their expertise to policymakers and others (e.g. the Integrity Institute). Given the pace at which a service's systems are updated and modified, the longer an expert has been away from working within a given platform, the more out-of-date their deep knowledge of its systems may be. And given the commercial sensitivity of the information at issue, both current and former staff are likely to be limited by non-disclosure agreements in what they can discuss about the technical inner workings of an online service.

Tensions and Tradeoffs

This, then, highlights a key tension in the quest to ensure that the requisite technical expertise is available to platform councils: independence versus pertinence. Platform councils need to trust that the expert technical advice they receive is accurate and not unduly influenced by financial, reputational, regulatory, or other considerations of the platform. However, the individuals with the most specific, detailed, and current understanding of how the platform's systems operate are highly likely to be people currently employed by the platform. And technical evaluations do not necessarily separate cleanly from other operational considerations: the feasibility of a particular recommended change or intervention in a system is likely to depend on considerations beyond just whether the change is technically possible.

A second tension is familiar from other discussions of platform accountability: transparency versus confidentiality. The work of a platform council will derive its legitimacy in part from the transparency of its operation and decisions. Transparent decision-making enables the affected parties and the outside world to understand the facts and rationale the council relied on to reach its decision. It also fosters additional learning and norm-development within the broader field of platform accountability. This includes transparency in the technical details of a case and information about how the underlying systems work.

Public details about the technical systems that underlie major online services remain relatively scarce and providing more information to the public is one of the many goals of the platform-council concept. It has already proven a useful output of the Meta Oversight Board, whose decisions have brought to

light new information about Facebook's "cross-check" program, "media matching bank", and machine learning classifiers for detecting nudity. Transparency in technical details also enables independent consideration of the tradeoffs involved in a given question, which can be useful for other online services grappling with similar issues as well as for researchers, regulators, and others trying to understand the online information environment as a whole.

In tension with this transparency, however, are the pressures towards keeping core technical information about an online service confidential. These pressures are many, including the legitimate need to maintain trade secret protection over certain intellectual property, the risk of manipulation of this information by bad actors seeking to abuse the service, and financial and regulatory concerns about the disclosure of derogatory information. As noted above, online services routinely require non-disclosure agreements from current and departing employees, which can significantly limit the ability of these individuals to share their specific knowledge of the operation of these systems. Platform council members may also have certain duties of confidentiality towards the services they evaluate (as well as the people whose claims they assess) that can further limit their ability to disclose information in published opinions.

The next section lays out a variety of options for bringing technical expertise into the deliberations of platform councils and discusses the tradeoffs between these various tensions.

Models for Incorporating Technical Expertise

There are multiple options for incorporating technical expertise into the deliberations of a platform council. A council will need technical advice and input at a variety of stages of its work:

- In reviewing and selecting the cases it hears, to understand the kinds of questions that are being put to it and to identify trends that may point to broader, systemic issues within the online service;
- In reviewing and evaluating the evidence and information provided by parties, especially that provided by the online service;
- In formulating questions and requests for further information from the online service; and
- In making a final decision and developing recommendations for how the online service should change its policies, practices, and design and use of different technologies.

In each of these phases (and likely others), council members will need to understand the technologies at issue, be able to ask thoughtful questions that identify how the operation of that technology may impinge on human rights, understand the information being provided by the online service and evaluate its legitimacy, and work towards technically feasible resolutions.

Appointing council members with relevant technical expertise

One way to ensure that relevant expertise is on-hand during the deliberations of a platform council is to appoint council members with a background in the kinds of technologies and processes the council will review. Depending on the size, composition, and remit of the council, this could include selecting council members with particular technical sub-specialties (much as one might appoint council members with expertise in the human rights law of different regions) or selecting members who have a general technical background and who can generally assess the information provided and ask relevant questions.

Positive aspects of this approach include ensuring that someone with technical expertise is present at every step of the process, and giving that person similar standing in the proceeding to other members of the council. Potential drawbacks include the risk of overreliance on one technical expert's perspective or understanding, particularly among non-experts. There is also a need to ensure a balance of expertise

across members of a council, to avoid either over-burdening a handful of technical experts or crowding out other types of expertise.

The pertinence versus independence trade-off may be particularly salient for the technical expert-as-council member; an individual with direct experience working on the systems of a given service will have the most pertinent expertise but may also face the strongest questions about the independence of their perspective from their former employer. These questions could, in turn, shape public perception of the independence and legitimacy of the entire council. Recent former employees may also face the most stringent confidentiality obligations, which could limit their ability to apply their particular expertise. It may be most useful to focus on appointing council members with more general expertise, e.g. from working on trust & safety issues across multiple online services, rather than designating specialists in each relevant service.

Receiving technical briefings from current platform employees

Current platform employees can provide council members with technical briefings on the general operation of the online service and on specific questions that arise in a given place. This approach is used by the Facebook Oversight Board, which describes a process in its charter for Board members to receive information from and ask questions of the staff of Facebook and Instagram about the cases on its docket.

These briefings can help with the council members' overall familiarity with the service's systems and processes and can provide opportunities for a more efficient exchange of questions and answers. They can also be helpful in building a relationship between the service provider and the council and fostering an environment of good-faith interaction. The downside is that such presentations will necessarily come from the point of view of the online service, which may not be self-critical or wholly forthcoming. Council members may not know the most incisive questions to ask, or be equipped to catch inconsistencies or over-generalizations in the explanations they receive from the online service.

In other words, this approach maximizes pertinence of the expertise at the expense of independence; the information will likely be exactly on point to the question under consideration, but it will lack independent verification. Online services may find it easier to reach confidentiality agreements to allow current employees to conduct these briefings, given their existing employment relationship with the staff members. But, especially given the lack of independence of this information, it would be essential for the council to be as publicly transparent as possible about the technical information that forms the basis of their decisions, in order to enable independent evaluation and critique.

Seek technical input from outside parties

Platform councils could embrace a role that exists in many court systems around the world: that of an *amicus curia*, or "friend of the court". Also called "intervenors" in many legal traditions, these amici are independent of the parties involved in a case and serve to highlight interests, interpretations of law, and factual information that may be useful to the court in reaching its decision. The Facebook Oversight Board employs a version of this function in the open call for public comment that accompanies each case announcement.

A council's openness to input from third parties can help socialize its work to a broader audience, potentially reinforcing the norm-development work of the council, and can yield unexpected information and perspectives to inform the council's work. It may be difficult to attract high-quality comments from third parties, however; in the case of the Facebook Oversight Board, a high-profile case may draw hundreds or thousands of comments (the case of Donald Trump's suspension received 9,666 comments)

while many cases see only a handful of contributions. There is also no guarantee that commenters will answer the questions the council members pose.

Third-party commenters may typically fall on the other end of the spectrum between independence and pertinence: they can be wholly independent of the online service being evaluated, but the information they have to provide may be general, out-dated, or off-topic for what the council really seeks to explore. One way of improving the utility of third-party commenter technical information is for the council to be as clear and specific in its call for comment about what it knows and understands already about the systems involved in the case, and to identify precise questions that amici may ask. Some third-party commenters may be limited in their comments by pre-existing confidentiality agreements (e.g. if they previously worked at an online service), but they will have considerable leeway in deciding how to share what information they can. Comments from third parties should generally be made publicly available.

Appoint technical amici to work with council members

Independent technical experts could also serve as special advisors to council members, as technical amici who engage in a sustained way in the deliberations around a case. Technical amici with appropriate skills and expertise could be identified at the beginning of a case, or even beforehand if the council intends to explore a particular set of questions about a service's systems. These experts could be on-hand to answer questions and provide general explanations about their take on the technical factors in a case and could be present at any briefings provided by the online service. The technical amici could also provide perspectives on the proposed solutions and recommendations of the council.

The availability of a technical amicus would allow for immediate clarification and discussion of technical information provided to the council throughout the deliberation and could enable deeper inquiry from the council and a swifter resolution of technical questions. It would ensure that multiple perspectives, not only the online service's, inform the assessment of the provider's systems. This structure could be vulnerable to over-reliance on a single expert's perspective, and the online service may not be as forthcoming with an independent third party involved in the discussions.

Appointing case-specific technical amici could provide councils with the flexibility to ensure that the individual's particular expertise is pertinent to each case. (As with every option for bringing technical expertise into council discussions, councils may find the pool of technical experts with service-specific insights limited by pre-existing non-disclosure agreements.) Concerns about the independence of a technical amicus with recent ties to the online service (whose advice would be highly pertinent) could be balanced by employing multiple technical amici, or by using an additional method for sourcing technical expertise. Technical amici would likely need to agree to some level of non-disclosure agreement to participate in case deliberations, in order to ensure robust disclosure from the online service provider. Once again, this necessary confidentiality can be balanced by transparency in the ultimate decision, with the council including information about the advice received from the technical amicus as well as the information provided by the online service.

Conclusion

Each of these models for incorporating technical expertise into platform council decision making involves tradeoffs between getting the most pertinent information, presumably from someone with ties to the online service being evaluated, and ensuring the independence of the perspective received from the technical expert. The need for access to specific information about the operation of a given platform's systems also raises issues of confidentiality, as online services may be reluctant to share certain information about their systems outside of the confines of non-disclosure agreements.

Ultimately, platform councils will likely need to use multiple sources of technical expertise, to account for the benefits and drawbacks of each approach. Technical briefings from the online service itself can be complemented by input from independent sources, including through public comments or an independent technical amici role, and at least some council members could be selected for their specific technical expertise and training. Whatever the source of the technical input, it will be vital for councils to explain their understanding of the systems and processes at issue, and how technical considerations inform their decisions and recommendations, as part of their published opinions. This will enable independent evaluation of the role of technical systems in various platform accountability questions and will help to increase the overall transparency of our information environment.

Further investigation into this issue should involve a deeper exploration of other oversight and decision making processes and their approaches to involving different types of expertise in their work. Media councils for print and broadcast journalism, for example, have been in operation in different parts of the world for many decades, and have served as the inspiration for today's debates about social media or platform councils. Deliberative democratic and collective dialogue processes seek to bring the people directly affected by governance decisions into the process of making those decisions; such processes must also grapple with the varying kinds of expertise that participants bring to deliberations, especially when considering technical questions. And regulators in many other industries, from aviation to pharmacology, routinely deal with highly technical questions.

Finally, it should be noted that platform councils are far from the only institutions who will be grappling with this need for enhanced technical understanding of the systems that make up online platforms. As legislators worldwide craft new intermediary liability laws and platform regulatory frameworks, judges and regulators will need to interpret and apply them, and will encounter many of these same questions. Platform councils, then, represent a useful opportunity to work through various options for incorporating technical expertise into third-party oversight of online services, and to balance these competing tensions of pertinence, independence, confidentiality, and transparency.

Enforcement as a necessity in platform councils

David Morar
OPEN TECHNOLOGY INSTITUTE, WASHINGTON D.C., USA

It is a worthy goal to democratize online communication spaces, but it will not be easy to achieve. By definition, any external institution built to infuse public values into the functioning of online platforms threatens the current make-up of the companies running those platforms, as they would have to cede the power they currently hold unilaterally. If platform councils are to perform this crucial task, one way or another, they would require enforcement capacity, or rather a way to ensure accountability.

Public values, rooted deeply in democracy, derive a significant amount of their legitimacy or worth from, among others, functions of accountability, as well as a right to rule, via consent and acceptance. For the purposes of this paper, these functions and rights are bundled together under the umbrella of enforcement.

If an institution inhabits a role of "governing" or "regulating" another or a specific group (or industry), that implies the existence of accountability, or enforcement, at a minimum through reporting, auditing, monitoring, and verification procedures. In the space of private governance, accountability measures range from being baked into the purpose of the institution, such as public assessments of compliance with a certain set of best practices, or certification schemes, all the way to exclusion from the institution. While necessary, these mechanisms are not sufficient to legitimize any institution, and even with stellar enforcement private governance structures may end up being ineffective due to a whole slew of other concerns.

Against this backdrop, this paper will first surface examples of critiques of similar institutions from other industries and their enforcement structures. It then looks at the most well-known example of a platform council, the Facebook Oversight Board, and explores its enforcement mechanisms. Finally, it problematizes the idea of the need for enforcement, by assessing industry perspectives on relinquishing its power and ways to alleviate concerns.

Critiques of similar institutions without strong enforcement

Multi-stakeholder governance institutions are institutions where power is distributed among actors from different stakeholder groups. These institutions have a very wide range of structures and varying degrees of enforcement/accountability mechanisms. Historically, those without strong, effective enforcement have been decried as marginally better if not similar to self-governance schemes, which is to say ineffective at allowing other stakeholder groups access to influence the actions of industry (whether individually or at industry level). Three critiques in particular can be very useful for envisioning enforcement within the platform council perspectives.

Blue/Greenwashing

The most damning response to multi-stakeholder initiatives is the accusation of greenwashing: to deflect blame for a misstep, a company creates or adopts a broad, substance-less set of principles which have no actual effect on the original issue. A similar version is that of bluewashing, where corporations join institutions or structures in functionally toothless labeling schemes where they benefit from participating alongside the UN, without making any meaningful changes to their social or environmental performance.

In cases when industry is not in any way beholden to making significant changes, a direct result of lack of enforcement capacity, multi-stakeholder initiatives are usually described by critics as a form of greenwashing. Businesses consent and accept only inasmuch as they want to be seen as doing so externally, but without any accountability, the institutions are considered toothless. The most famous example of bluewashing is the Global Compact, which is predicated on companies adhering to nine principles related to environmental protection, human rights and labor standards. However, it has very few ways of actually enforcing this adherence.

Designed/heavily influenced by businesses

Another criticism of multi-stakeholder institutions is that industry's role in creating, implementing, or running the institution overpowers any kind of significant input by others, usually civil society. Beyond having the power to actually influence or make the rules, industry would also have the power to dilute the accountability aspect of the institution. While providing its consent and acceptance, the company stakeholder group becomes the one in charge of its own accountability, either directly by wielding power within the institution, or indirectly, by building the structure themselves.

For instance, industry-led institutions that claim to be multi-stakeholder and multi-stakeholder institutions with a strong if not overpowering industry role are truly a difference without a distinction. This critique is is usually leveraged against multi-stakeholder institutions regardless of their actual make-up, so identifying real examples is crucial The Sustainable Forestry Initiative is an industry-led standards body that claims to have equal representation of stakeholders, but its output is far less stringent than the also flawed but multi-stakeholder Forestry Stewardship Council, which makes its true nature apparent.

Lack of potential to grow stronger

A more nuanced critique of multi-stakeholderism is that its powers rarely expand beyond its original scope. It's important to remember that multi-stakeholder institutions usually exist because companies have ceded some of their own power. As previous critiques argue, what is presented as a legitimate power of the institution is, in fact, not real, or hobbled by industry (alone or with others), so the hope would be that with time the institution would amass real influence. However, an institution built with limited enforcement can't, by itself, decide to increase its own enforcement capability unless the entity being governed further agrees to it. What potentially can happen is the decrease of the enforcement capacity and thus authority of the institution.

Internet governance has famously been experimenting with multistakeholderism. One particular institution of the internet governance ecosystem, the multistakeholder Internet Governance Forum (IGF) was a peculiar outgrowth of power struggles between the previous multilateral UN-based regime and the upstart multistakeholder community taking part in the ICANN (Internet Corporation for Assigned Names and Numbers) which de facto governed the infrastructure layer of the internet. The outcome was bizarre: after a failed attempt at wrestling power from ICANN, in 2005 the UN ended up with a compromise that its newly created Internet Governance Forum would not actually have any outputs, thus effectively neutering the organization, and subsequently removing any kind of enforcement. The IGF remains an institution whose main goal is to bring stakeholders together, without any kind of policy output.

While such general critiques are important in the abstract, far more useful is to understand how one of the few platform councils in existence fares on these issues.

Meta Oversight Board

The Meta Oversight Board's structure shows that it was built with a focus on deliberation, rather than governance. The enforcement mechanism and thus the legitimacy of the Board is limited in both scope and dimensions. Designed by and for a corporate entity, with the ostensible and amorphous goal of oversight, the Board has become a beacon of primarily externalizing responsibility for difficult disparate choices instead of a democratizing check on industry power.

Providing its deliberative action with a contractual mandatory accountability role signifies that the company has to implement, within the boundaries of predetermined rules, a decision by the Board's panel. While no penalties exist for non-compliance, it would certainly lead the entire edifice to crumble. The legitimacy of the Board has as one of its many necessary but not sufficient attributes the concept of the enforceability of its judicial decisions; Facebook consents and accepts these decisions a priori.

Its role was always first and foremost to provide a limited and pointed check on specific thorny cases of content issues. Thus, it was invested with contractual power to provide enforceable decisions in those cases. Given a perfunctory and wholly performative power to comment on matters of policy, the Board fails to inhabit a role of meaningfully inserting public values in its relationship with Facebook. Facebook does not automatically consent or accept the comments from the Board that are outside of the judicial cases. Even more, there are no explicit or implicit accountability mechanisms, as evidenced in the 2021 X-Check case, where the Board learned from media reports that Facebook lied about one of its programs, and subsequently issued an advisory opinion in December 2022 with no way of reprimanding the company.

Confined to its limited role, FBOB was properly designed as an external adjudication body. Extrapolated to the discourse surrounding it as a governance or rather "oversight" structure, the FBOB was faulty, to say the least. It can be argued that a platform council with a limited remit and with even more limited mechanisms for enforcement, starting with a right to rule and ending with accountability, would have a hard time actually infusing public values into the corporate space. Certainly its existence by itself can denote a shift towards democratic values, as a small portion of the company's power is diminished and re-routed this way. However, the important decisions related to the functioning of the platform are still taken by the business. The FBOB's founding documents, built by the company, which was relinquishing power, did not provide for a way to actively embody and wield legitimate power over the overall functioning of the company.

However, adding such mechanisms to a platform council can potentially lead to the exact opposite scenario, where companies feel they do not have a way to respond.

Enforcement critique

A crucial aspect of decision making at any level is the enforceability of its outputs. A decision that can be ignored without consequence undermines the structure. At the level of the state, Max Weber famously posited that a monopoly of legitimate use of force would be necessary to enforce order. However, the state, and especially a democratic state, has throughout its mechanisms important checks and balances. Be it, like in the US government, where the three branches serve as each other's enforcement mechanisms, or between citizens and the government, where those who have consented and accepted to be governed have the opportunity to use their own power to hold those in power accountable.

Whereas a democratic nation-state's duty is to secure basic rights for its citizens, a corporation's incentive is to uphold its duty to its shareholders. Thus, within democratic nation-states the enforcement

mechanism is, at least ostensibly, wielded by entities with similar incentives to those being overseen, ultimately securing said basic rights. Not only are they not similar, but platform councils would be the product of the company relinquishing its power to a group of people who are not aligned in their incentives. Even more, why would a company give up their power to an institution which it has no way of reprimanding, or at least calling out potential concerns? Arguing for a strong enforcement mechanism would make it even less palatable to corporations to relinquish their influence over their own products and services.

The counter to such an argument should be in the details of the organization of the councils. Building effective enforcement would also require ways for those being governed to object, and do so in a way that allows for legitimate complaints, but does not burden or grind to a halt the structure. Including as equal input the desires of the companies, or allowing them to have a voice in the process would be a way to do so. For instance, the Internet Corporation for Assigned Names and Numbers, ICANN, has a robust commercial stakeholder group with a nominally equal voting bloc in matters of policy.

Conclusions

When attempting to understand an experiment like platform councils, it is useful to learn lessons not just from the limited examples in the field but also from similar cases from other industries.

This paper has three major takeaways. First, built-in mechanisms for consent, acceptance and accountability, from clear contractual terms (in the case of ICANN) all the way to dissociation (in cases like the FSC) are important and they can be one of the deciding factors in whether the infusion of public values is effective. Second, as certain actors (most likely industry) can take advantage of a weak structure or of their own power, institutions built on decoupling the powerful (in this case platforms) from their power require enforcement mechanisms to ensure actual governance or oversight and to avoid pitfalls of perceived or real legitimacy gaps. Third, platform councils that do not include in some legitimate way either the voice of the platform in its deliberations or mechanisms for complaints may find it difficult to implement enforcement mechanisms.

Interoperable Platform Democracy: How deliberative democratic processes commissioned by corporations can interact with nation-state, multilateral, and multistakeholder decision-making

Aviv Ovadya
BERKMAN KLEIN CENTER, CAMBRIDGE, USA

Is there a world where corporations not only run democratic processes for their decision-making—but where that process is actually a *good* thing? A world where important and controversial choices facing corporate platforms and AI organizations are decided not by leadership fiat but by a truly representative deliberation (largely outside of government)—and where this is not just 'democracy washing'?

This piece explores what that world might look like, and how such democratic processes—potentially commissioned by corporations—might beneficially interoperate into our existing institutions of national, transnational, and global governance (hereafter referred to simply as institutions).

Such questions are particularly salient given both Meta's concrete actions to use such processes (initially to develop greenfield policies in the Metaverse), and AI leaders' exhortations to "align their interests to that of humanity"—where such processes might be particularly applicable.

There were several key guiding questions which led to the approach outlined below:

– Who *should* actually be in charge?
– Is it possible to govern tech in a way that moves power to the people being impacted—and away from *both* corporate leadership and oppressive governments?
– What are **pragmatic** approaches we can try **today** to rapidly improve the governance of transnational technologies—in a world where such international coordination seems increasingly difficult?

What is platform democracy?

Platform democracy: Governance of the people, by the people, for the people—except within the context of an internet platform (e.g. Facebook, YouTube, or TikTok) instead of a physical nation.

More formally, platform democracy refers to the use of democratic processes to include the populations impacted by a platform, in the governance of that platform in a representative fashion.
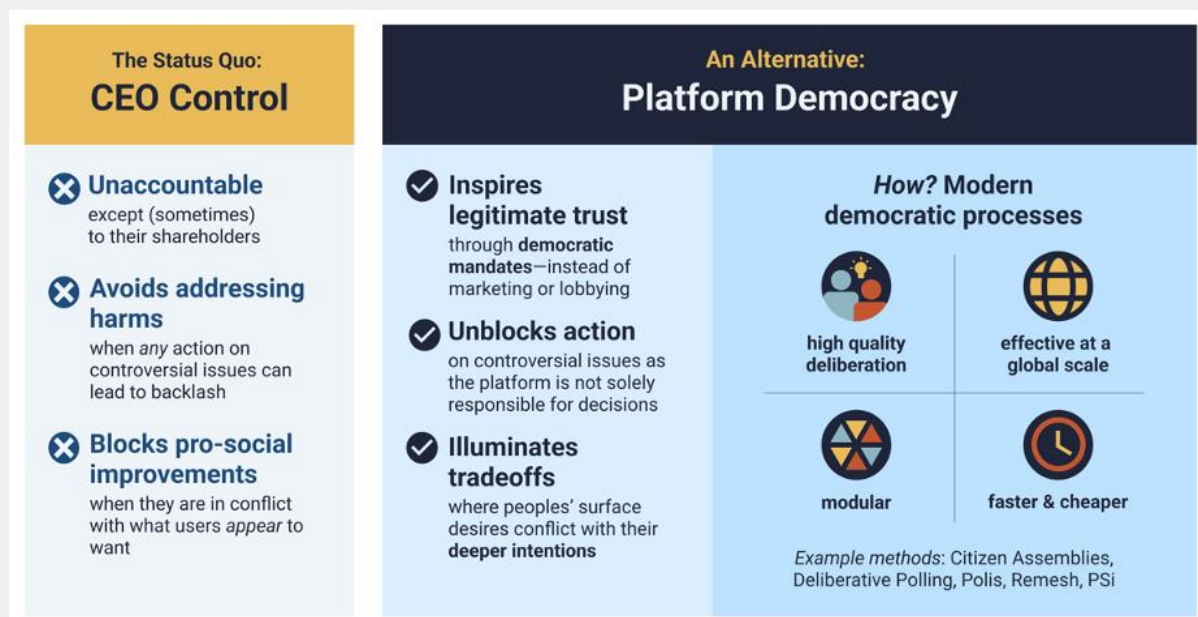
In particular, two approaches to such platform democracy are considered here: intensive deliberative democratic platform assembly processes for complex decisions and lighter-weight collective dialogue processes for decisions that need less context. In both cases, an organization (such as a platform) needs to make a decision that would benefit from democratic legitimacy from the decision.

Such questions might include:

- What if anything should be done about content that is not strictly false, but which is meant to be misleading?
- Under what conditions, if any, should audio or video be recorded in online spaces in order to identify potential harassment, and if so, who should have access to such recordings?
- What kinds of content, if any, should not be shown as 'trending'?
- What kinds of outputs are acceptable from generative AI systems?

None of these are theoretical. Meta has already directly explored a version of the first two of these questions through such processes; Twitter would likely have asked the third question had there not been an acquisition, and OpenAI's CEO has described the fourth as a question for which he would like global democratic input.

How are these questions then answered such that the processes are "democratic"?

A microcosm of the impacted population is convened and facilitated by a neutral 3rd party, such that everyone being impacted by the decision (might be e.g. the user base, or the countries the organization operates in) has roughly the same opportunity to be selected (through sortition: stratified random sampling). The selected people make the ultimate recommendation to the decision-maker—and unlike a poll, they are given the opportunity to learn from each other's perspectives (and for decisions that involve significant tradeoffs or context, they also learn from stakeholders and experts). These 'deliberators' are paid for their time, and ideally child care, elder care, travel, etc., to reduce the self-selection bias.

Why are these processes legitimate?

The potential democratic legitimacy of such processes comes from their representative nature—instead of every single person having an opportunity to vote, but spending fairly small amounts of time per person, a much smaller number of people vote, but they each are supported with the time and resources to make the best possible decision (without the often perverse incentives of electoral politics or corporate profit). Such processes are also not just some techno-optimistic idealistic dream but are being used by existing governments around the world. Moreover, as any individual has only a small chance of being selected, it is far more feasible to imagine such processes working across many platforms, even globally, than an electoral representation system.

Can deliberative processes work across many languages and cultures?

Such deliberative democratic processes have now been run with many languages a number of times, including several across the EU. Admittedly, interlingual and intercultural deliberation is still imperfect, but there are both process approaches and tools that can help mitigate the risks, and ongoing experimentation to develop best practices around the most challenging aspects (e.g., subtle differences in word connotations across languages).

What happens after the process is complete?

As a slight generalization, when governments run such deliberative processes, they usually serve as recommendations, which must be either implemented, or receive a response from the government about why the recommendation is not being followed. The same can apply when the commissioning organizations are companies like Meta or OpenAI; though it is also likely possible to make the results binding.

Interoperability with existing institutions

Platform democracy does not exist in isolation—it should be structured to support existing institutions instead of fighting them. There are several places where this can happen.

To contextualize the options for interoperability with existing institutions, it's useful to understand the different organizations potentially involved in such a process.

- First, there must be a **commissioning organization**, e.g. Meta, Twitter, Google, or OpenAI. This could also be a combination of organizations, or even organizations and governments together.
- They commission the deliberation with "**deliberative infrastructure providers**"— organizations that run these sorts of processes as neutral third parties for governments (and now companies) around the world. The deliberative infrastructure providers also select the members of the deliberation body using sortition and facilitate the assembly itself.
- These deliberation infrastructure providers may work with **existing expert and stakeholder bodies** to provide context for the deliberators or create a **new temporary advisory or stakeholder body** to help support the deliberation.
- Finally, the **deliberation members** learn from those stakeholders, experts, and each other in order to make the final recommendation.

Commissioning Organisation: For example, social media, search generative AI org.

Deliberative Infrastructure Provider: For example, citizen assembly facilitator, collective response system

Expert & Stakeholder Body: For example, an existing multistakeholder body or one-off expert and stakeholder groups created for the assembly

Deliberation Members: Sortition selected members of the public, forming a representative microcosm of population impacted by a decision

Impacts of platform democracy outputs

The most obvious touchpoint where such a process interacts with the broader world beyond the commissioning organization relates to the impacts of the process *outputs*.

Media pressure

Let's not beat around the bush here—platform democracy, in its most limited form, can be considered a form of self-regulation. However, it is different from most forms of self-regulation in that the power of creating the mandate is not directly in the hands of the platform. It is instead put to people chosen at random, without any incentive to accede to the platform's wishes, and facilitated by a 3rd party deliberation organization. (A rigorous process aiming for strong legitimacy would also use as impartial a method as possible of choosing experts and stakeholders.)

Moreover, even if the recommendation is not binding, the legitimacy of the mandate created by such a representative democratic process makes this kind of self-regulation rather awkward for a company to ignore. In plainer words, it looks very bad to the media (and the governments that follow it) if an organization convened a process to have the people tell it what they want in a democratic fashion—and the organization ignored the outputs.

Perhaps even more exciting, a sufficiently transparent and high-profile process not only educates the members of the deliberation themselves, but allows the broader public, media, and even regulators to follow along through broadcast and social media, enabling learning from experts and stakeholders alongside the members. This can potentially help elevate the overall level of conversation on issues with complex trade-offs more effectively than hearings used to score political points, and also help the public see itself through its reflection in the deliberative microcosm.

Raising the responsibility baseline

In fact, recommendations that come out of a process that is seen as broadly legitimate are likely to not only affect the organization that is convening it, but also any other organization facing similar questions and advocacy and interest groups that relate to the question (assuming it is not too specific to the convening organization). If the question is for example around potential responsibility actions, this can help create a corresponding *responsibility baseline*—a minimal level of action that is seen to be broadly acceptable, which may be higher than the current industry default, raising pressure to implement responsibility practices across the board.

Even if the responsibility baseline is lowered, that is potentially indicative that the impacted population does not actually believe that that level of 'responsibility' is warranted (for example, you can imagine a deliberative process that determines that there should actually be less content moderation around a particular issue—which would be a good thing to know).

Creating a responsibility 'north star'

Some processes may not change the baseline, but may instead create a north star— responsibility practices that might be too difficult to fully execute on, but can be aspired to and approximated. Such north stars may also exert pressure on the entire industry of the commissioning organization.

Identifying global 'moral high ground'

For some issues, the challenge is not around the ideal north star, or the minimal baseline for responsibility. Instead, there might be deeply competing notions of what responsibility even is. For example, some organizations developing powerful AI systems say that the responsible thing to do is to share as much as possible—maximizing openness. Others are extremely cautious and barely release any information about their research. Both sides say that they are acting for the good of humanity—in other words that they have the 'moral high ground'. Both argue their perspective to the public and regulators with the intent of shaping perception and law. Similar differences in approach occur in many domains, including in tradeoffs between privacy and security.

A rigorous global deliberative process can create something closer to an idealized public sphere to actually identify what 'humanity' believes is the moral high ground (that such companies should be aiming for). There are thus potentially strong incentives for organizations that believe that they are closer to the 'true moral high ground of humanity' to convene such processes, in order to have their approach validated (assuming that they are correct).

Regulatory and institutional suggestions

Such responsibility baselines, north stars, and moral high grounds may then directly impact the actions of legislators, regulators, standards bodies, multilateral bodies, multi-stakeholder bodies, trade associations, etc., in ways that may be binding. In other words, the commissioning organization is essentially fronting the cost for deep research and input gathering that can then directly feed into these existing processes, some of which may have more binding force. Concretely, this might look like, for example, the UK government, the EU, UNESCO, or the Partnership on AI developing recommendations (or, for governments, even laws) directly based on and referencing those deliberative outputs. This could be true even if the deliberative process that was originally convened by Meta or OpenAI—assuming that the process was seen as rigorously impartial and democratic.

Bindingness

There is the option that the convening organization can pre-commit to making an output binding (when otherwise legal), using the legal infrastructure of the jurisdiction(s) they are operating in. There are likely a number of legal instruments that can be used to do this depending on the relevant jurisdiction (e.g. a golden share arrangement).

Conflict with platform democracy outputs

There are some commons questions about how this might play out in practice:

What happens when there is conflict with existing law or regulation?

In situations where there are conflicts between the outputs of deliberations and existing law or regulations, the situation is roughly analogous to when a company's strong ideological stance conflicts with that of a government. In some cases, this may be seen as good, e.g. when a company avoids sharing location information about democracy activists, thus violating the laws of an authoritarian country. In other cases this may be seen as problematic, e.g. when a ride sharing company ignores local safety regulations. Either way, if organizations do not follow the laws of the nations they are based in, they face the consequences. The main difference is that if the legitimacy of the process used to create the deliberate outputs is higher than that used by the government (for example in an authoritarian or extremely partisan context), then there may be significant pressure, both externally and internally pushing for the more democratic outcome.

Could the governments or regulators themselves actually be involved in the process?

Definitely, though of course this can become more challenging with more global processes (and thus more governments). It's also worth noting that one of the benefits of the platform itself running a process is that the process can be specific to features that only that platform has, and it may not be worth the time for government officials to be involved with every platform in such a manner. That said, especially for processes which involve multiple platforms or industry consortia, governments may want to act as co-convenors, and platforms may want that also in order to increase legitimacy of the outcomes.

Could there be permanent deliberative bodies?

There are many potential models beyond the simple temporary platform assembly or collective dialogue, including institutionalized permanent models built on approaches such as multibody sortition, the

Ostbelgien model, the Paris model, and which could directly interact with existing institutions in much more sophisticated ways. It feels somewhat presumptuous to explore this in the context of platforms and companies without more understanding and exploration with the basic model, but it is important to know that it may be possible to have key decisive power over an entire company through such processes, as they are refined and combined. One could even imagine augmenting or replacing a traditional corporate board structure with carefully designed deliberative bodies in order to truly enable democratic governance, with no higher executive or board level power (though feasibility might depend on the jurisdiction).

What happens if there are multiple representative deliberations with conflicting outcomes, perhaps even some run by the governments themselves?

There is no clear answer to this as this entire regime is too nascent. It is perhaps roughly analogous to having multiple treaties or non-binding agreements which are in conflict in a multilateral context. The ideal is likely that the process which is most rigorous and thus most legitimately democratic wins out—but there are many potential interpretations of rigorous, legitimate, and democratic, and no clear arbiter. This suggests that it is particularly important to create international standards for such processes in order to ensure consistent evaluation.

More generally, any time there are multiple competing decision-makers, potentially of varying quality, and no official hierarchy, there is bound to be tension, (ideally productive tension) and there is value in creating institutions to navigate those tensions.

Inputs to platform democracy

Beyond simply interoperating with other organizations through the outputs, democratic processes *inputs* can also interact with existing institutions and organizations at other stages of the process.

These include:

- The *commissioning organization* could actually be a joint body involving a partnership of a platform (or platforms) with a government or even governments, multilateral institutions etc. The commissioning organization could itself be an existing multi-constituency body such as the Digital Trust and Safety Partnership.
- The *expert and stakeholder body* could also be an existing multi-stakeholder body such as the Partnership on AI.
- Governments could help support the actual *process of sortition selection* if they already have 'sortition infrastructure' (as e.g. Mongolia has, as illustrated by its incredible turnout for their deliberative democratic process on constitutional amendments).

Why we might want platform democracy

I might prefer a world where purely public institutions fully govern our technological developments and have kept up with the rate of technological change—change that respects no borders. But we have not evolved our existing governance institutions to take on the challenge of legislating at the speed of technology, and that is unlikely to change very quickly.

The realist question we thus face is:

"How can we practically govern an onslaught of technological disruption—and what are the consequences if we fail to do so?"

This is not theoretical—platforms like Facebook, YouTube, and TikTok have shaped society through their policies, but even more impactfully, they have shaped the incentives of society through their ranking systems. These ranking systems determine what kinds of politicians, journalists, or entertainers succeed and shape the kinds of content they produce. Our existing governance institutions, over a decade after this became clear, have done very little to improve the impact of such systems on society outside of the narrow scopes of personalization and privacy.

We can and must do better, both to tackle belated issues and the emerging governance challenges around new technologies. This is especially salient for advances in AI, such as foundation models like GPT-4 and the products like ChatGPT built on top of them, which are likely to rapidly transform our lives. Perhaps deliberative democracy can help us find a way forward. Given a steady rhythm of convened processes, decisions might be made democratically, even at global scale, within months instead of years or decades.

Platform democracy alone cannot solve our problems, but it perhaps provides a useful new governance option between our status quo of platform autocracy and platform chaos.

Public Values and the Private Internet

Peter Routhier
INTERNET ARCHIVE, SAN FRANCISCO, USA

Introduction

The Platform://Democracy project asks how we might "ensure public values in private and hybrid communication orders," and what role social media councils could play in achieving this result. This paper takes as its starting point that the relevant "communication order" is the internet itself—not just the "walled gardens" of today's social media companies. From this standpoint, we can see a few ways social media councils could inadvertently hinder the development of public values in the broader online ecosystem. And we can suggest one way to address this: by enabling our existing public interest institutions, like libraries, to perform their traditional functions in the online environment. In this way, we can build public values into the broader internet.

Social Media Councils and the Broader Internet

For many years now, Tim Berners-Lee and other early internet pioneers have identified certain aspects of social media as presenting "threats" to the World Wide Web—and the internet beyond the Web itself. Others have been more blunt in their assessment: "The internet has a Facebook problem, but the internet is not Facebook." While often critical, these kinds of voices also typically express optimism about our ability to make the Web whatever we want it to be: "We create the Web . . . We choose what properties we want it to have and not have."

From this standpoint, we should consider what impact social media councils could have not only on today's dominant social media platforms, but on the broader development of the internet and the World Wide Web. This paper suggests three possibilities: they could unduly legitimize the already-existing social media platforms; they could entrench these platforms into their dominant positions; and they could distract from the potential to build public interest values into a broader and more diverse online ecosystem.

First, social media councils might unduly legitimize today's social media corporations. These companies are largely duty-bound to maximize their profits, and there is reason to doubt that they are willing or able to make this duty subservient to the advice of a council or otherwise. If one accepts that at least some of the problems with social media companies that councils are seeking to solve stem from the profit-maximizing behavior of those companies vis-à-vis the operation of their platforms, then—under Delaware law, at least—the influence of a social media council will be muted if it is not able to supersede or at least somehow coexist with the profit-seeking directives at the heart of their corporate governance. In these circumstances, social media councils might lend to platforms the imprimatur of their members and their organizers, all while being structurally incapable of imbuing them with their values.

Second, social media councils can work to entrench today's platforms into their dominant positions. If these corporations are unwilling or unable to voluntarily alter their problematic behavior, as the first point holds, then the natural conclusion of the regulator is to make mandatory and superseding rules. But in a variety of ways—most notably, the cost of compliance—such rules can entrench these companies in a position of dominance online. That is, imposing new compliance obligations on platforms of all stripes can have counterproductive structural consequences. It is possible, for instance, that they would increase the cost of participation in the relevant online environment to such a degree

that only corporations with a certain level of financial resources can bear the costs. This could increase, rather than reduce, the profit-maximizing behaviors of the relevant firms which led to the initial problem to be solved. And it could remove the possibility that alternative organizational models could arise or persist in the new environment.

Finally, social media councils can distract from the potential for different kinds of reform. The best example here may be Facebook's creation of the limited liability corporation that it calls the Oversight Board. The process leading to the development of the Oversight Board corporation was no doubt fascinating and extraordinary: the company has spent hundreds of millions of dollars on its creation and development, capturing the attention and interest of a variety of actors from around the world. Meanwhile, the Oversight Board corporation itself has written only a handful of guidance documents with which it requests Facebook's voluntary compliance. Has this amounted to much more than a distraction? Was it intended to be anything but?

None of this is to say that social media councils cannot do real good—it would seem to be the least we could ask of some of today's dominant platforms—but rather to suggest that we shouldn't let them legitimize our existing platforms, entrench them, or distract us from other possibilities for imbuing public interest values into our online information ecosystem. One such possibility would be to empower existing public interest institutions, like libraries, to play a larger role—even simply their traditional roles—in the digital environment.

An Internet with Public Interest Values

When considering how we might imbue the broader internet with public interest values, a natural starting point would be our traditional public interest institutions, like libraries. To be sure, social media companies have displaced an older model of information production, publication, and curation that was populated by a variety of different actors and constructed in a fundamentally different way; we cannot "put the genie back in the bottle." But it is worth considering whether and to what extent we could empower public institutions like libraries so that they can, at least, play their traditional role in the digital environment.

Libraries are relevant to the discussion for many reasons; they have long played "a fundamental role in society" as "gateways to knowledge and culture." They have done so by, among other things, preserving and lending books for the public's benefit. Importantly, unlike social media companies and a variety of other actors to come before them, libraries often have a public character rather than a private, profit-maximizing interest.

As a result, it is perhaps no surprise that some have called for librarians to play a more active role in the online information ecosystem. But as with other experiments with placing traditional knowledge workers into social media environments, placing libraries and librarians directly into the social media framework may offer no salvation. Indeed, there is reason to believe it may not be desired by many librarians themselves. In 2022, Internet Archive and the Movement for a Better Internet convened a group of leading librarians and others at Georgetown Law Center for a workshop on digital library issues. As the resulting whitepaper reports, those present expressed little interest in working as fact checkers or content moderators for the Facebooks of the world. What they hoped for, instead, was to be empowered—or, at the very least, simply allowed—to fully execute their traditional public interest functions in the digital environment. Unfortunately, as outlined in the paper, a variety of economic and legal factors have so far worked against their doing so. But as Ethan Zuckerman has noted, we should not assume that the current "economic and legal system govern[ing] our online spaces" will persist in perpetuity; such thinking "obscures possible solutions to the challenges arising around the socially corrosive effects of new media technologies."

For an example of how empowering libraries in the digital environment could improve our online ecosystem, consider Internet Archive's project to weave books into the web—and turn all references on Wikipedia blue. In short, this project has made hundreds of thousands of authoritative sources—books, webpages, and more—available to readers and contributors of Wikipedia. Vast quantities of Wikipedia citations are now cite-checkable, so that readers and contributors alike can check their sources and work within a more complete and accurate information environment.

Conclusion

While many have proposed structural reforms which would weaken the power of dominant social media companies or otherwise open them up to competition and reform, we should also consider empowering public-interest-minded actors so that they can play a larger role in the online information ecosystem. Social media councils may help bring public values into private companies, but we should not stop there. Empowering public interest organizations like libraries, so they can execute on their traditional functions in the online information environment and seek to innovate towards new ones, can help bring public values to the broader internet as well.

Soft Power and Platform Democracy: How social media councils could shape government and corporate strategies and preferences

Fabro Steibel
INSTITUTE FOR TECHNOLOGY AND SOCIETY, RIO DE JANEIRO, BRASIL

In previous work, we have addressed the question of content moderation and accountability as well as how to use policy framing to understand how institutions are designed to solve policy challenges. The previous research showed how fact-checkers and judiciary institutions have framed the problem of negative advertising (the terminology used to refer to fake news back in 2011) in largely different ways. The relevance is related to accountability: if we set different policy questions to solve, we have different methods to evaluate what solution is ideal to cope with content moderation.

In this article we emphasize using policy framing to explore if the concept of platform democracy is going in a similar direction. Are social media councils being framed in largely different ways? And if not, how do the policy framings used to describe the values, policies, and institutions related to social media councils help us to understand where the content moderation agenda is going?

Platform democracies is a concept that explores the role of social media councils as external governance structures. The councils are more commonly initiated by companies themselves, as a self-regulation practice based on external oversight. One of the main cases is the Facebook Oversight Board, but many other cases are worth noticing, from Ireland's planned "Social Media Council" to TikTok's "western" formed council.

Platform democracies are far from a stable model of regulation, but they foster intense debates around the limits of co-regulatory models, including unusual models of "regulated self-regulation". Brazil is one of the countries with advanced legislative debate. According to the model, legislation mandates platforms to set up councils (as self-regulation bodies), prescribe them with content moderation and procedural rules, and supervise them with a state-controlled agent. The idea has been presented in other contexts, including the German policy debate.

But how can we understand the innovations in policy framing driven by these "platform democracies" debates? This is the overall question we pose in this short article.

Soft power refers to "the ability to affect others through the co-optive means of framing the agenda, persuading, and eliciting positive attraction in order to obtain preferred outcomes." Soft power focuses not on changing a subject's strategy (what hard power does), but rather on curtailing a subject's agenda, its first preferences. Hence the relationship between soft power and behavioral change, or constructivist theories such as framing, agenda-setting, and priming.

What is soft power?

Soft power is a term coined by Joseph Nye that has been widely applied as a resource for public diplomacy and state actors. For instance, Naren Chitty notes that soft power "as a term, represents a body of thought that is associated with resources invested in attraction-power as well as with strategies for using such resources to further actors' interests." The underlying idea is that while, in general,

governments remain the most powerful actors on the global stage, the stage is increasingly more crowded.

Critics of the terminology argue that other concepts such as ideology or culture are best fit to frame a subject's power preferences. Supporters of the terminology are aware of such criticism and argue that the concept fills a void in global international policy that helps to explain how governments (and large corporations) act.

In practical terms, soft power can be used for example to persuade individuals to join large groups, to cooperate in resource distribution, to facilitate political organization or even to frame prescribed actions. In this way, soft power is less a matter of governing and rather a matter of cooperation, learning, and growth, especially in larger contexts.

The elements of soft power in platform democracies

To measure how soft power influences platform democracy, we must rely on categories that are mostly intangible, such as culture, values, political ideals, and institutions. In this particular case, we decided to focus on three specific variables, namely values (e.g. human rights and democracy), policies (e.g. obligations to remove certain contents), and personalities (e.g. actors or institutions involved in platform democracy). The source of data (in particular the emphasis on variables used to frame the policy debate) is based on overall debate, but it is possible to evaluate in quali-quanti terms the reference to terms used in policy discourse.

Values

Three key values are frequently used to justify the policy framing of platform democracy: individual rights and human rights, digital constitutionalism (which encompasses a number of other values, including the rule of law, separation of powers, among others), and tech exceptionalism.

The values related to individual and/or human rights are a direct consequence of the selection of sources used to explain platform democracy. Voices and context are mostly selected from the EU and US regions, rather than coming from China and Russia, for example. This is true even for social media councils from companies based outside western regions, such as TikTok.

With this in mind, it is easier to understand the emphasis on fair procedures and individual rights protection (a reference in particular to the EU region) or from corporate social responsibility principles, such as external oversight (a reference in particular to the US context). In the same direction, platforms and governments in the debate emphasize the valorization of Human Rights standards, in particular those grounded in the United Nations Guiding Principles for Business and Human Rights (UNGP), framed as a "soft law" that "defines a corporate responsibility to respect human rights."

A second value is the rule of law, a cornerstone of the digital constitutionalism narrative. Looking at it this way, we can understand the relevance of state-like principles in the private domain, as in the case of platforms assuming state-like obligations usually attributed to state authority. Hence the importance of values such as procedural fairness, transparency, decision-making accountability, access to information, participatory governance structures, and more.

Lastly, we consider the value of "tech exceptionalism," the value that some companies are not traditional companies, but rather a special type of company. On the one hand, this value highlights that excessive regulation could stifle innovation and limit economic growth, which can be avoided with a more flexible and adaptive regulatory approach. On the other hand, the value focuses on the particularities of content moderation activities by online intermediaries, leading to debates around algorithmic curation, content

removal and agent behavior. The problem here is one of scope: in the digital ecosystem, social media platforms do not work mainly with online content intermediation.

Tech exceptionalism ends up setting a division between digital companies, one that may not have ground in real markets. In this policy framing value, platforms are not digital companies per se (because they work with content-moderation while others don't), but platforms may also profit from other digital markers (such as the link between digital payments and social media).

Policies

We identify three key policies frequently used in the overall debate to justify the policy framing of platform democracy: terms of service, multistakeholder governance, and co-regulation.

Rules and guidelines regulating social media platforms are usually codified as "terms of service," a set of legal agreements between the provider and its users that outlines the rules, responsibilities, and guidelines that users must adhere to when using the platform's services. Those policies highlight for example how content is taken down or stays up on online platforms and are increasingly used by governments to justify actions where state regulation is yet to be adopted, with a potential hybridization between state and non-state governance.

Multistakeholder governance refers to the policies used by platform democracy settings to institutionalize decision-making models that promote participation from non-state and non-private actors, notably those coming from academia and civil society organizations. The focus on multistakeholderism promotes governance models that move away from classic industry-led self-regulatory organizations, and away from state-driven hard regulation, favoring a 'hybrid governance' system.

Lastly, we identify the emphasis on co-regulation. The policy frame here favors that platform democracy is best fit to avoid the pitfalls of both hard regulation and self-regulation. The innovation comes in the form of variations of the co-regulatory model, that lead to formats that emulate platform democracy as advisory, quasi-legislative, and quasi-judicial institutions.

Institutions

When looking at the plethora of institutions involved in platform democracy, we identify an emphasis on voices coming from journalists, civil society, academia, and platforms.

Fact-checkers, and journalists in general, are often believed to be mandatory actors to better understand mis/disinformation content-moderation. This group is portrayed in the overall debate as individuals who should be supported (including with economic incentives), and who have a role in social media councils to contribute for example as suppliers of evidence or as community voices.

Civil society and academia are emphasized in the overall debate as highly valuable members of social media councils, occupying roles related not only to advice, but also on rulemaking and decision-making. There is also emphasis on diversity of members composing the boards, a selection criterion set from start when creating such forms of external oversight.

Lastly, we highlight the role of platforms as institutions. Social media councils are mostly defined as non-binding forums of decision-making for platforms, which highlights the role of companies themselves to participate in the debate, as high-level voices, and decision-makers.

What next on framing platform democracy's soft power?

It seems likely that a consensus will emerge around the pros and cons of social media councils as content moderation oversight bodies. The level of consistency between values, policies and institutions shows very little variation indeed.

Priming values explain why platform democracy values focus on individual rights, human rights, and digital constitutionalism. The association favors the expansion of state-like regulation to private companies activities. Some read this new frontier as problematic, making private-sector potentially arbitrary and contradictory decisions as validated state-like mechanisms. Others read the scenario in a positive light, considering such councils as an innovation to overcome the worst uses of private sector or state sector for speech moderation.

Another crucial value is explaining social media companies within a tech exceptionalism light. Some will point to the risks of such an approach in overlooking market concentration aspects; others will see this aso the necessary justification to advance in new models of co-regulation even further.

It is also important how the participation of civil society and academia is defined as mandatory. Some will see in this case an opportunity for better oversight of platform self-regulation, while others will consider the participation of such actors as insufficient.

A Council for Consilience: How could a council foster a field researching the information environment?

Alicia Wanless
CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE, WASHINGTON D.C., USA

Councils have long played a role in human history, providing an organising function to bring some community to bear on a problem. Today, democracies face a pressing and complex problem in a polluted information environment. Yet the very act of governmental intervention to address challenges within the information environment raises questions about democratic legitimacy. Many desired interventions, such as banning bad actors and disinformation from social media platforms, resemble authoritarian approaches and are being implemented simultaneously as trust in public institutions plummets. Moreover, little is known about the impact of many of these interventions because research of this type is just emerging, and its practitioners lack consilience and are not supported to do such work at scale and over time. Thus, we have a situation where policymakers need to understand the information environment which is largely controlled by private businesses, and researchers currently study in silos from various disciplines without common terms and methods. In democracies, this means that any attempt to do something about the information environment must be inherently multistakeholder, even if that first step is simply to understand that system. Could a council approach enable a multistakeholder effort to better understand the information environment in the context of democracy?

What's the problem?

Much of the problem is that we don't know what we don't know. The information environment isn't being studied as a system, making it difficult to put research into context. It is tempting to try to assess the effects of specific pieces of technology, individual threats like disinformation, or campaigns run by threat actors, but if we don't understand the system in which they operate, it is impossible to understand the effect of one variable. The two major reasons for this ignorance about the information environment are a lack of consilience, or shared understanding across disciplines, and gaps in resources supporting research.

The consilience problem stems from the fact that researchers assess the information environment through the lens of their own disciplines, whether in media studies, ethnography, computer science, or psychology. This leads to a lack of sharesd terminology and methods for studying the information environment. It's difficult to study a system if we aren't seeing its entirety and building on the work of others in a systematised manner. For a scientific understanding of the information environment to emerge, a convergence of disciplines must occur, with researchers speaking a shared language and working consistently together on the topic.

The second issue of resource gaps is wide-ranging. Gaps include researchers' lack of access to data generated by social media companies. Researchers also lack access to platforms to conduct studies, such as measuring the impact of interventions. Even if researchers had access to social media data, many lack the engineering resources and infrastructure to make sense of that data at scale. This is to say nothing of

a lack of diversity among researchers who mostly come from North America and Europe and, even then, do not represent the diverse communities and experiences of all those living there.

In short, massive gaps impede an understanding of the information environment at the very time that democratic societies must quickly get a handle on the role of that system, lest the interventions introduced today have unintended consequences.

It takes a village...

Filling these gaps requires a multistakeholder community. Stakeholders from across sectors must be brought to bear to fill the many resource and research gaps. There are four key types of stakeholders: researchers from academia and civil society who lead in framing understanding of the information environment; governments who set policies and enforce regulations related to the governance of the information environment; companies that often control major aspects of the information environment; and citizens whose agency to make free and informed decisions within the information environment is necessary to democracy and therefore should be a part of understanding and governing it as well.

Fostering consilience entails connecting the work and approaches of different fields to allow researchers to build on each other's work and collectively develop an understanding of how the information environment works as a system. It requires systematising what is already known about the information environment, finding consensus on shared terminology, and building frameworks for studying the information environment that can transcend time, geographies, and the inevitable introduction of new technology that changes how we process and communicate information. While efforts are emerging to systematise existing research, for example, through an Intergovernmental Panel on Information Environment, other questions like how the information environment should be framed and studied as a system and what other types of data can be used to do so also need answering.

Researchers have also been working to address the resource gaps outlined above. The European Digital Media Observatory Working Group on Platform-to-Researcher Data Access is developing detailed guidelines for researcher data access. The University of Michigan is developing the Social Media Archive, accepting data deposits, and making more data available for research purposes. Social Science and Humanities Open Cloud for the Netherlands will help researchers share data from images, surveys and social media. The Observatory for Online Human and Platform Behavior at Northeastern University "captures the online behavior of a large sample of volunteers" generating data for research. Princeton University has begun developing shared engineering infrastructure. These initiatives should be commended, but if democratic societies want to understand the information environment at the speed necessary to address current challenges, this community must come together to foster consilience and champion shared resources at a scale that can support a larger community of researchers than in single countries or their own institutions.

Beyond researchers, governments are also an important stakeholder. Democratic governments have an obligation to understand the impact of the interventions they are making in the information environment, lest the cure is worse than the disease. Governments can affect significant change that leads to an understanding of the information environment faster through funding to fill resource gaps, and also by mandating that online services publish operational reporting to increase transparency on how they operate and share data with researchers.

Tech companies are also key. They bear a considerable responsibility to support research on the information environment, given their role within and the revenues derived from it. Regulation will also

increasingly compel industry to find solutions to safeguard their users and ensure their operations are not degrading the information environment.

Another stakeholder that is frequently the focus of interventions but often forgotten in consultations is citizens. Finding ways to meaningfully engage them in research on the information environment is key for ensuring the longevity of democracy. Citizens should absolutely be informing the ethical principles that govern research of the information environment, not to mention the lifecycle of data generated by them through social media.

While much activity is underway across stakeholder types, these efforts are often disconnected from (and sometimes at odds with) each other. The paper now considers whether councils offer a method of efficiently convening multiple stakeholders without condemning efforts to elite capture, particularly by governments and industry.

…or a village council

Councils have been used by a variety of communities to solve problems. A council – the Council of Europe – was used to bring a fractured continent together after the ravages of war. But they have also shaped and controlled religions. Revolutionaries saw councils as a means to overthrow the existing order. Others pinned hopes for greater democracy on citizen engagement through councils. More recently, the concept has been applied to hold tech companies to account through social media councils. In short, councils are what people make them. So, what makes a council?

Councils represent a sort of community, coming together to collectively make decisions like solving a problem. Councils tend to be framed by rules governing their activity, persist over time, and comprise a limited number of selected people. For councils to emerge, they need a catalyst – they take work. These limitations can impede the level of diversity of council members, and of democracy, most councils can expect to employ. In other words, councils embody a paradox of top-down action to foster a bottom-up approach. What follows is an exploration, drawing on the author's current efforts to build a council in ways that leave room for a community to emerge in a ground-up fashion as members develop it together.

In this example, Jacob N. Shapiro and I are starting with an idea that other researchers have also expressed: creating the equivalent of a European Centre for Nuclear Research, a CERN for the information environment. This multinational research facility, the Institute for Research on the Information Environment (IRIE), would develop shared infrastructure to speed up research on the information environment to inform policymaking. A council guiding IRIE's development might form for a year or so to answer challenging questions that would inform the institute's development and operations.

Resist the urge to answer everything

I have never built a multinational, multistakeholder research centre before. Few probably have. I can tell you that our instinct, perhaps as North Americans, was to get a group together and build it. But as a small number of us puzzled through the next steps, the difficult questions needing answers piled up. Like, what ethical principles should govern this research or what funding models will ensure the independence of IRIE? At the time of this writing, fifteen questions have emerged that directly relate to how something like IRIE should be structured. There are likely more. Many of these questions are challenging and given the relationship of the information environment to democracy, they shouldn't just be answered by a small group of researchers from elite institutions in one part of the world. But even if

we wanted to answer all these questions ourselves, it would take years of research. My first lesson has thus been to resist the urge to try to have all the answers at the outset. This problem is immense. But if a wider research community comes together, as part of a council, to answer these questions, we can speed up research to support policymaking.

Start Small

Trying to build a council from the ground up can seem counterintuitive. The trick is constructing a tent that can remain open with enough space to welcome new and unforeseen community members as they emerge. With that in mind, maybe the council isn't designed to be completed with a fixed number of members and launched fully formed. Rather, it should develop over time based on the number of questions that arise needing to be answered before IRIE could ever operate. Council building could start smaller by identifying who might already be answering those questions identified or whose work might lead to answering them. This can be done through consultations with other stakeholders working on related issues. Through this process, a list can be generated to begin mapping the wider community. In some consultations, leaders might emerge who volunteer to take on questions. In other cases, the work of someone might be so advanced it makes sense to approach those leaders and gauge their interest in being part of a growing community working towards a common goal of building a field and providing it with shared infrastructure.

Keeping diversity in mind as the community forms is important, but we are all limited by our own networks that introduce a certain degree of exclusivity by virtue of whom we know. Moreover, my limited and privileged experience might preclude me from even conceiving of the barriers to entry others might face. Indeed, the way some of these questions are framed might already be introducing bias that makes it hard for researchers to engage from other countries and backgrounds than mine. To that end, it isn't sufficient to simply construct a council with a mix of backgrounds over geographies but to include questions as they arise from researchers with a diversity of perspectives. Likewise, it isn't enough to invite someone to a table under our own terms, but we must make space for those who come later to inform our collective efforts, quite possibly from a vantage point we haven't yet considered and help ensure that these insights are heard by others.

IRIE COUNCIL QUESTIONS

- What can science already tell us about the information environment?
- How should the information environment be framed and studied as a system?
- What kinds of data can be used to study the information environment?
- How will legal regimes governing aspects of the information environment affect IRIE?
- What is the mechanism by which data can be made available for research purposes?
- How can citizens be better engaged in research on the information environment?
- How can a greater diversity of research be fostered?
- What is the relationship of data storage to the research lifecycle?
- What are the most pressing research questions that need to be answered, and how can they be studied?
- What ethical principles should govern this research?
- What funding models will ensure the independence of IRIE?
- What kinds of large-scale engineering infrastructure can speed up that research?
- How can a feedback loop be best created such that research informs policy-making?

– How can capacity be developed to help researchers make use of IRIE and new opportunities to data access offered by the Digital Services Act (DSA)?
– What is the impact of this research on the physical environment?

This approach fosters a community to answer the above questions and work together towards a shared vision. While some community leaders might already be working to answer integral questions and be well-supported, others might need to be identified to start work and will need funding. The key to success is taking a collaborative approach to work together and matchmake interested donors with community members willing to undertake answering pressing questions while also being open to growing the council as new questions need to be answered. In answering their questions, researchers might be encouraged to take a mixed-method approach and engage as broad a community as possible to encourage diversity. Ultimately, it is up to that leader to decide how best to answer their question and whom to engage in so doing. Each leader would be asked to develop a short one-to-two-page proposal outlining their plans and resource requirements. Taken together, these proposals feed into the requirements for a wider council that can be fundraised together or in parts, with many donors engaged across geographies to support the emerging effort.

Wait for a Home

While most councils start with a home and are built top-down starting within, building from the ground up might mean waiting to find a place to house it. In this case, it involved putting an idea out to the community, bringing an initial group of committed leaders to bear, and stepping aside to allow that emerging community to choose its leader and a home for the future council in a third-party organisation that can carry the project forward.

For a multinational research facility like IRIE, it's important that such a home be non-partisan and not viewed by the wider research community as competition. If the aim is to speed up research on the information environment to generate evidence for policymaking, then it must be accessible to the widest community of researchers possible, facilitating their research instead of competing with them. It must foster collaboration. Philanthropies with a long track record of supporting or recognizing intellectual endeavours, like the Nobel Foundation and Kofi Annan Foundation, come to mind. Ultimately, the initial members should choose where this emerging council is housed.

Find Funding Carefully

It is up to leaders to choose funding sources that fit their comfort level. The least complicated sources usually are philanthropies. For the perceived integrity of the overall project, who funds the council itself and provides it with a home must be chosen very carefully and not be affiliated with politics. Donors who support leaders answering research questions can be more diverse, opening the door to include smaller philanthropies in this work. This approach also enables democratic governments to directly support researchers in their country, contributing to the overall effort.

Something at the scale of a multinational research facility can only be sustained over time following a similar approach to the CERN, whereby multiple democratic countries agree to contribute each year based on their GDP. To protect IRIE from perceived interference by donors, this commitment must be made over a long-term basis into decades and not up for political approval on a renewal basis. While few countries have the resources to fund such a centre alone, thirty countries sharing the burden of contributing $1M yearly start to make a multinational research facility a reality. However, given the timelines most governments work on, it could take years before such commitments are made. Taking a

ground-up approach to building and funding the council piecemeal helps the community collaborate now to have the answers needed when that day comes.

In the short term, drawing on industry money to answer IRIE questions will be discouraged. No matter how well-intentioned that support might be, the perception of industry support could taint the overall project. In the long term, one of the key questions that must be answered through this process is how IRIE can be sustainably funded and maintain its independence. It would be unreasonable and perhaps unjust not to expect social media companies and those who profited from them to support efforts to understand and improve their effects, but all parties must recognize the need to protect against the reality or appearance of research being shaped by donor interests.

Hinder Capture by Industry and Government

Unfortunately, many tech companies and governments are not without past transgressions when it comes to the information environment. Both stakeholder types also wield more power than those in civil society. For these reasons, researchers from across academia and think tanks bear the burden of leading on answering questions and engaging stakeholders from industry or government in a manner that can prevent capture by the latter. However, stakeholders from multilateral organisations representing democratic governments could potentially lead in answering questions, as their mandate to support member states often entails information gathering rather than shaping the agenda of an external effort to suit the needs of a specific country.

Build a Council

A council might begin to emerge after a certain percentage of the identified questions have leaders committed to answer them. Yet, to encourage diversity and enable a wider community to become engaged, this might be a point to stop and try other forms of outreach, such as hosting a meeting on the side of a related conference, such as the International Conference on Computational Social Science or the Paris Peace Forum, depending on the community needing to be reached. Indeed, finding opportunities to bring leaders and the wider community concerned with the integrity of the information environment together creates a means for growth and fostering a field. It reinforces the purpose of existing conferences while also fostering community-building more efficiently. After all, the information environment is complex, and the problems are so numerous and challenging that there is more than enough work for everyone. We must work together. Moreover, through these engagements, more leaders will emerge willing to answer the remaining questions or new ones as they emerge.

The council is ultimately formed of those leaders answering questions in pursuit of a shared aim, such as building a multinational research facility to study the information environment. Together, the council chooses a chair and a home. The function of the council in the short term is to share information between those answering different questions to ensure each member's efforts help head in their shared direction. In some cases, answering one question, such as how the information environment should be framed and studied, might feed into the answering of another, for example, what data might be used to study it and vice versa. Each leader must publish a final report as part of their effort.

While the initial purpose of the council is to share information between members to help the community work towards a shared aim, as a final step, it could also be tasked with drafting the plan to build a multinational research facility. Indeed, each of the final reports would feed into this process. How the institute would be governed and structured would ultimately be determined by this process, informing

what role a future council might have, if any. And in this manner, a community can develop to tackle a complex issue more quickly.

Keep the Tent Open

Even with a council formed, it must not be viewed as a fait accompli. Questions that had not been considered will inevitably arise, and answers needed to inform plans for a multinational research facility. Council members should resist the urge to answer more questions themselves and engage others who might emerge later wanting to support the initiative. They can support other leaders but should lead on only one question. This will help build community and collaboration to address more outstanding questions faster. Similarly, others might emerge who want to contribute shared infrastructure or research help in finding a path to keep the tent open to newcomers so as to support as wide and diverse a community as possible across democracies. Indeed, several of the known questions relate to keeping the tent open, such as in finding ways to engage citizens in research on the information environment and fostering greater diversity of research.

Newcomers might fit two categories: those who take on answering key questions that inform the governance and structure of IRIE and those who contribute to practical operations. The former would consist of leaders tackling a new question that must be answered before a plan can be drafted for IRIE and should be incorporated into the council. The latter are core long-term partners for IRIE should it come to exist in whatever form it takes. This could include philanthropies interested in supporting the effort, policymakers with pressing problems that need research to address, as well as a broad range of actors from the research community, such as those making data available to study, those conducting measurements on aspects of the information environment, others building tools to support such work, and the wider community who would benefit from shared infrastructure. Essentially, the shared vision is the tent under which we build and foster a wider community committed to understanding the information environment in the context of democracies.

Conclusion

There are many ways to build a council, each with its pros and cons. As noted, the approach outlined in this paper is paradoxical; building a council from the bottom up still requires top-down action to start. It is an awkward chicken-and-egg situation. Not taking on the entire leadership to build a council can make the process unruly and not intuitive for others to follow. Likewise, funding in pieces to support the answering of individual questions might not be compelling for researchers to want to step forward and tackle problem sets, but on the flip side, it enables greater outreach to the community and the ability to engage a variety of donors and form networks to support future work.

Building a council from the ground up is inductive and, as such, must be open to continuous improvement. Indeed, existing council members should remain open to new findings. This doesn't necessarily mean changing course entirely, but being able to adapt as we all learn together. To that end, it should have an inherent feedback loop between council members enabling their findings to help inform the work of the whole and, ultimately IRIE. We all bring something to the table, and making sense of the information environment will require many perspectives and skills. Consilience will take compromise. We are working towards the same end. Instead of saying no, we could commit to finding solutions. Together these three ideas could form the principles framing the council.

If this approach works, a community working together can do more and faster. Likewise, working together as part of a wider community could, in turn, create a groundswell across democracies to

champion IRIE and make it a reality – something none of us can do alone. Governments in democracies won't tackle something this big if researchers in their own backyards don't see a need for it. Moreover, while not intuitive, keeping the tent open could help identify a wide network of researchers, each studying aspects of the information environment from their own lens to help build the consilience and diversity in research that is so badly needed.

Only time will tell whether this approach will work or not. Although, at the time of writing this, of the fifteen known questions, researchers have stepped forward to answer nine. Even if the bigger goal of creating IRIE doesn't happen, the act of community building in pursuit of it could foster a field, and that alone would be huge.
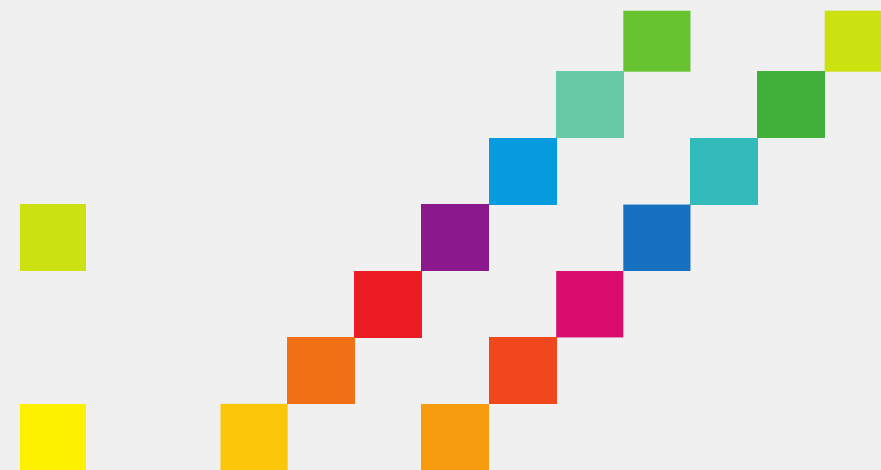
MATTHIAS C. KETTEMANN, JOSEFA FRANCKE, CHRISTINA DINAR AND LENA HINRICHS (EDS.)

Platform://Democracy

Research Report Asia-Pacific

PLATFORM://DEMOCRACY

Platform://Democracy

Perspectives on Platform Power, Public Values and the Potential of Social Media Councils: Research Report Asia-Pacific

edited by Matthias C. Kettemann, Josefa Francke, Christina Dinar and Lena Hinrichs
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

Contributors

Name(s)	Affiliation
Setu Bandh Upadhyay	Technology Policy and Digital Rights Analyst
Bayan Koshravi	Mehralborz Higher Institute of Education, Tehran, Iran
Lucinda Nelson	Queensland University of Technology, Brisbane, Australia
Prateek Waghre	Internet Freedom Foundation, New Delhi, India

Table of Contents

Introduction to the Asia–Pacific Research Clinic

Setu Bandh Upadhyay
TECHNOLOGY POLICY AND DIGITAL RIGHTS ANALYST

The Asia-Pacific research clinic aimed at examining the normative infrastructure needed to establish better rulemaking, rule-enforcing, and rule-adjudication structures in hybrid communication spaces, primarily online. Specifically, the clinic sought to explore the potential of platform councils as a solution to address legitimacy deficits in private and hybrid orders of platforms, which are characterized by idiosyncratic normative logics, vocabularies, and algorithmic arrangements. The clinic also examined the decentralization of power and alignment of public values with private orders, while exploring various governance models for social media platforms.

The clinic focused on the Asia-Pacific region and brought together a diverse set of perspectives. Discussions delved into nuanced topics such as the impact of social media councils when dealing with different languages, fact-checking, Web 3.0, communication infrastructure, local platforms, value-sensitive design, legitimacy of self-regulation, and transparency. The clinic's approach encouraged the fellows to engage in short, but deep diving into specific regional challenges instead of going for broad overviews. In this way, each short paper was able to contribute meaningfully to the debate, and created some appetite for more deep dives into regional avenues.

The clinic was composed of experts in social media governance, journalism, social justice, human rights, law, technology policy, digital media research, and platform governance. The fellows represented the Asia-Pacific region, including Iran, India, Sri Lanka, Singapore, Hong Kong, Taiwan, and Australia. Additionally, the participants had expertise in other jurisdictions such as China, East Asia, and South Asia.

Themes and Outputs

Governments can encourage or 'nudge' the adoption of private ordering through platform councils as a way to improve the platform ecosystems. Civil society, scholars, and activists are advised to closely monitor the level of market concentration and the rate of innovation in platform ecosystems. This is to ensure that the benefits of platform councils outweigh any risks associated with it. For example, platform councils may lead to increased competition and innovation, but it could also result in the exclusion of certain groups or the exploitation of vulnerable individuals. Policymakers may consider exempting smaller platforms and open knowledge community platforms from major platform regulations and compliance requirements that are meant for larger platforms.

Regarding specific platforms with regional challenges, one study explored the role of social media as the main channel of communication in schools and how value-oriented and democratic design can be leveraged to improve social media governance.

Another study focused on regional case studies in an effort to study how cultures and norms of acceptable behavior are shaped on social media platforms. Using feminist theory, the study showed that Social Media Councils can publicly expose gender inequality in content moderation by calling attention to instances where content moderation policies or practices unfairly target or harm women or other gender minorities and that they play a crucial role in shaping rules relating to gender equality through their interpretations of vague policies in specific cases.

One study criticized the executive-appointed social media council in India with a focus on their formation, legitimacy, capacity, scale, and expertise. The study highlighted that an executive-appointed

and controlled council is not desirable due to concerns over low levels of operational transparency and vagueness of remit across multiple bodies and intermediaries.

The Impact of Private Ordering on Platform Competition

Setu Bandh Upadhyay
TECHNOLOGY POLICY AND DIGITAL RIGHTS ANALYST

Key Findings

This paper explores how private ordering is linked to competition. The paper uses theoretical analysis to establish linkages due to a lack of regional examples. The paper concludes with recommendations that talk about nudging by regulators/government for platforms to adopt private ordering, encouraging a deeper exploration of the relationship between private ordering and competition, and suggests private ordering as a solution to facilitate a competitive platform ecosystem.

Introduction

Social media platforms like Meta, Twitter, and YouTube have become vital to the global information landscape. How they manage user-generated content has significant implications for freedom of expression, privacy, and public discourse. This is where private ordering comes into the picture. Platforms like Facebook and Instagram already have the Oversight Board to regulate their content policies and decisions. And Twitter also had a Trust & Safety council of independent experts that was dissolved by Elon Musk. Spotify, a music platform with a social feature, also has a safety council. While this has provided a degree of democratic voice in the governance and regulation of platforms, it has also raised concerns about censorship, bias, and the influence of dominant players in the industry.[1]

The advantage of private ordering is that it allows for a more flexible approach to regulation. Businesses and industries are often better equipped to understand their operations' unique challenges and requirements and can take proactive steps to self-regulate. For example, platforms can respond to emerging trends in online harassment, hate speech, and misinformation by implementing new policies and tools to address these issues. This can result in a more effective and efficient response than government regulation, which often moves at a slower pace. However, private ordering by social media platforms can also have drawbacks, such as a lack of transparency, accountability, and consistency. The platforms may not always be transparent about their content moderation policies and may enforce these policies in inconsistent or discriminatory ways. In some cases, platforms may act to limit free speech, suppress dissent, or otherwise undermine democratic values and human rights.[2]

Private ordering can also create barriers to entry for new competitors, limit consumer choice, and entrench dominant players. The impact of private ordering on platform competition depends on the specific design of the private agreements and the broader regulatory and market context in which they

---

[1] Newton, C. (2022). To build trust, platforms should try a little democracy. Retrieved February 14, 2023, from Platformer at https://www.platformer.news/p/to-build-trust-platforms-should-try.

[2] Kettemann and Fertmann (2022). Platform-proofing Democracy—Social Media Councils as Tools to Increase the Public Accountability of Online Platforms. HIIG. Retrieved February 13, 2023, from https://www.hiig.de/publication/platform-proofing-democracy-social-media-councils-as-tools-to-increase-the-public-accountability-of-online-platforms/.

operate. As platforms move towards private ordering, it is crucial to understand this trend's impact on competition and the overall regulatory ecosystem in the context of platforms.

If platforms have a competitive advantage when they use platform councils, it may suggest that private ordering can potentially lead to market concentration by players enacting platform councils resulting in an overall reduction in competition. On the other hand, if platforms do not have a competitive advantage when they use platform councils, it may suggest that private ordering can promote competition and innovation. In either case, private ordering can play a role in promoting democratic control over markets to redistribute ownership and control of communication infrastructure by providing a voice on governance to people who are not only driven by profits.

Private Ordering and Competition

Empirical studies suggest that competition among platforms often responds in unanticipated and sometimes ambiguous market positions.[3] This is clearer when it comes to platforms that work on user-generated content – i.e., social media platforms. The research reveals that consumers' content preferences are governed by network effects, potentially indicating a correlation towards their intention to have a voice in content and platform governance.[4]

Industry standards set by private ordering can improve the efficiency and effectiveness of operations and increase trust between the industry and consumers. While platform councils can provide a flexible and adaptive approach to content moderation challenges, they can also create barriers to the entry of new players, community-generated or hosted platforms, and customer choice in the way of additional financial burden or increased internal compliance in jurisdictions where they face no regulatory burden.

Wikipedia has been a great example of providing a democratic and collective voice in the governance of a significant platform, now known as the "Wiki-Governance" model.[5] In 2022, The Indian Ministry of Electronics and Information Technology requested that the Wikimedia Foundation furnish details on their response to the defacement of Wikipedia pages relating to two individuals named Arshdeep Singh - an Indian cricketer and an Indian footballer.[6] This request was made in accordance with the Information Technology Rules, 2021, specifically sections 3(1)(b) and 3(1)(d), which prohibit intermediaries from hosting, storing, or publishing illegal information that poses a threat to the sovereignty and integrity of India, among other things.[7] Subsequently, the affected pages were placed under "semi-protection" to limit editing access to trusted users. The entire incident, including the adaptability Wikimedia could rely on using their moderators, demonstrates the benefits of private ordering, industry norms, and user control over the platform, which also seems to have no adverse effect on the competition.

Private ordering can have a significant impact on competition and innovation. On the one hand, private ordering can encourage competition by promoting uniformity in business practices and allowing for a level playing field. On the other hand, private ordering can also limit competition by creating new business entry barriers and stifling innovation. It can lead to an unfair competitive edge for dominant

---

[3] Cennamo, C., & Santalo, J. (2013). Platform competition: Strategic trade-offs in platform markets. Strategic management journal, 34(11), 1331-1350.

[4] Zhang, K., & Sarvary, M. (2011). Social media competition: Differentiation with user generated content. Marketing Science, 47, 48.

[5] Dove, E. S., Joly, Y., & Knoppers, B. M. (2012). Power to the people: a wiki-governance model for biobanks. Genome Biology, 13, 1-8.

[6] Agrawal, A. (2022). No, India didn't 'summon' Wikipedia over Arshdeep edits. It asked for information within 24 hours. Newslaundry. https://www.newslaundry.com/2022/09/06/no-india-didnt-summon-wikipedia-over-arshdeep-edits-it-asked-for-information-within-24-hours.

[7] PRS Legislative Research. (2022). The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. PRS Legislative Research. https://prsindia.org/billtrack/the-information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021.

players in an industry. When a few large companies control a large share of the market, much like the social media landscape, they may be able to use private ordering as another tool to entrench their position and limit the entry of new players.

This can lead to market concentration and anti-competitive behavior. In such cases, private ordering can stifle innovation and limit the ability of smaller, less established platforms to enter the digital sphere and sustain. This ends up reducing the number of choices available to consumers. One famous example of undemocratic private ordering which directly led to stifling competition has been the Motion Picture Association. The association used the film ratings system as an unreasonable restraint on trade to smaller studios and independent films.[8] This comparison with social media platforms makes sense because similar to MPA, there are few prominent players making decisions on what content may be viewed by users and not.

Moreover, if Platform Councils are not transparent and open to only participants who are driven by profit (e.g., the company appointed majority), it can lead to a lack of oversight and accountability, which can again result in anticompetitive behavior in the form of collusion and harm to consumers. When used responsibly and with oversight, Platform Councils can lead to more efficient, democratic, and effective regulation.[9]

Measuring the impact of private ordering on competition and innovation can be challenging, as it can be difficult to determine the precise effects of private ordering on competition policy in the digital age. Digital markets are have many crucial variables which can influence competition, such as multisided markets, powerful network effects, economies of scope/scale, large amounts of user data, disruptive innovations, integrations, and switching costs.[10] For these reasons, analyzing competition in digital markets and assessing whether changes are needed to existing competition policy frameworks with respect to private ordering needs in depth study of said features.

However, it is possible to assess the impact of private ordering on competitiveness by examining the level of market concentration, the rate of innovation, and the level of consumer choice. Although private ordering by social media companies can have positive effects on competition and innovation, it is important to carefully monitor the process to ensure that it is fair, transparent, and open to all participants. This will help to promote healthy competition, foster innovation, and ultimately benefit consumers.

---

[8] Kilburn, C. E. (2013). An Offer You Can't Refuse: A Sherman Act Antitrust Examination of the Motion Picture Association of America and the Use of the Ratings System as an Unreasonable Restraint on Trade. UMKC L. Rev., 82, 255.

[9] Klonick, K. (2020). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression (SSRN Scholarly Paper No. 3639234). https://papers.ssrn.com/abstract=3639234.

[10] OECD (2022). OECD Handbook on Competition Policy in the Digital Age. https://www.oecd.org/daf/competition-policy-in-the-digital-age.

Recommendations

This piece is only able to cover selected aspects of private ordering and competition and ends with a call on competition scholars to analyze the impact of private ordering on innovation and competitiveness on platforms. As discussed above, private ordering offers several benefits, including quick adaptation, increased innovation, and personalization. However, on close inspection, it is revealed that private ordering also carries risks, including lack of oversight, anti-competitive behavior, and market concentration.

Therefore, it is important to monitor, explore and study the relationship private ordering has on specific industries, including platform businesses. The findings have important implications for stakeholders. Policymakers must consider the impact of government intervention on competition and innovation and the potential consequences of relying too heavily on private ordering.

The following key recommendations are proposed:

5. Platform councils and other forms of private ordering improve platform ecosystems while ensuring a fair, democratic, and competitive environment.
6. Civil society, scholars, and activists should pay careful attention to the potential of private ordering to increase or decrease the level of market concentration and the rate of innovation to ensure that the benefits of private ordering are greater than the risks.
7. Policymakers in Asia, particularly in places where technology policy and platform regulation are in nascent stages, may want to take inspiration from European Union's Digital Services Act. Smaller platforms and open knowledge community platforms should be exempt from major platform regulations and compliances.

Conclusion

Private ordering has the potential to significantly impact innovation and competitiveness both positively and negatively. It is, therefore, vital to carefully consider the benefits, risks, and challenges that come with private ordering in comparison to government intervention. By taking a balanced approach, policymakers can promote a healthy business environment that supports competition and innovation.

It is important for governments, civil society, activists, and other stakeholders to engage with social media platforms to promote a healthy balance between private ordering and public oversight. Platform councils may help deal with improving transparency, promoting accountability, and encouraging the development of best practices and industry standards. By working together through platform councils, regulators, platforms, users, and other stakeholders can help to promote a healthy and vibrant online environment that supports freedom of expression, privacy, and public discourse while also giving breathing room for a competitive environment and alternative platforms

Which role can Social Media Councils play in educational contexts? The case of the Shaad Platform in Iran

Bayan Koshravi
MEHRALBORZ HIGHER EDUCATION INSTITUTE, TEHRAN, IRAN

Abstract

I build on the findings of a case study to illustrate how Social Media Councils (SMCs) can have an effective role in an educational environment under conditions of high government interventions. Even though the Shaad platform was almost entirely regulated by government agencies, I argue that SMCs could play a significant role in the enforcement of rule of law and appeal mechanisms. I also argue that, in similar contexts, the proper design of SMCs must consider inclusion as the primary criteria for ensuring that enforcement and appeal decisions reflect democratic values such as diversity and participation. Lessons learned from this case could be useful in forming more inclusive SMCs.

Case Background: Shaad as a social media platform

On the eve of the Covid-19 lockdown and school closures in Iran, the Iranian government introduced Shaad as a communication and educational social network that was to provide the platform for teaching in public schools across the country for the next two years. This platform was a substitution for all teaching, communication and administrative tools that previously were available in person. It was developed by Iran's largest telecom operator on the basis of a general-purpose Iranian social network called Rubica. Shaad was offered free of charge and users incurred no cost for accessing content or data usage. Three roles were defined for users: students, teachers and headmasters. Nearly all Iranian students were automatically registered on the network, and it is estimated that about 70% of Iranian students used it daily. Despite the increased burden on parents to be involved in their children's education, no monitoring or supervisory role was created for parents on the platform.

While much of the content moderation services were provided by third parties (mainly teachers), the Ministry of Education of Iran made all defining decisions about how the system should be used in schools and what types of educational content can be provided on the platform. Teachers' primary responsibility was to moderate the content posted by students and to ensure that content used for educational purposes met the standards set by the ministry. There are also reports that suggest that certain groups of inspectors were set up in every province to monitor whether schools used Shaad. This issue was important to the government since it wanted to be sure that all students have access to educational content while schools were closed.

Establishing a social media council in this context could have a huge impact on the operation and further development of the Shaad platform. At the very least, there would have been ample opportunity for involving parents, teachers and other third parties like educational technologists in the regulatory process. There are parents-teachers associations in approximately every school in the country which meet regularly to discuss issues related to schools and the education of children. No arguments have been made against replicating this system of parents-teachers associations (or a modified version of it) on the platform. It seems that the opportunity simply escaped the minds of policy makers and parents

were far too unfamiliar with the platform and too concerned about the consequences of the Covid-19 pandemic to demand involvement.

Even though such councils were never formed, Shaad could benefit from the use of an SMC in an educational context. Some issues that an effective educational SMC can solve include safety and privacy issues, barriers in communication on the platform between different actors who should be involved in the educational process of students, and the presentation of inappropriate content.

These governing issues require a multifaceted approach that goes beyond government regulations. While regulations can set a standard for online behavior and safety, they are often too broad and generalized to provide effective solutions to the specific challenges faced by educational social platforms. An effective social media council is essential in providing tailored solutions to above-mentioned issues by taking into consideration the unique needs and values of the platform's users. Therefore, the diversity of expertise and ideas of SMC members will provide a holistic approach to governance that ensures the safety and well-being of users while promoting productive, engaging and meaningful educational experiences. Furthermore, an effective SMC has the flexibility to adapt and respond to changing trends and emerging challenges as they happen.

Case Analysis and Discussion

Since Shaad was limited to an educational context and was developed and operated wholly inside of the country, there were no political concerns about the flow of information that might challenge the government's authority. There was also no legal framework for the operation of such platforms. Activities on Shaad were only subject to broader guidelines concerning the educational system and online protection of children. Much of the government concerns were directed at dealing with online bullying, ensuring access to age-appropriate content and adherence to educational standards. None of these topics are politically controversial as they are nearly universally accepted. Therefore, a possible role of SMCs could be targeted towards applying and enforcing these rules and values on the platform. In this way, SMCs can be useful to address the problem of content governance. This role is now mostly limited to teachers whereas there is a high potential of using the capacity of other stakeholders.

However, this does not suggest that the design and operation of such councils would be an easy task. Iran is a vast country with diverse cultures and a complex ethnic composition. Thus, interpretation and adherence to regulation require local adaptation in ways that cannot be predetermined by central administration. Therefore, SMCs in similar contexts should be based on promoting multistakeholderism within the rule of law at a local level. Since multistakeholderism is often aimed at ensuring the inclusion and equal participation of stakeholders in the collective decision-making process, in this section, I discuss the principles that must underpin the design of such efforts in similar contexts.

Identity Management and User Training are Critical for Ensuring Inclusion in SMCs

Inclusion in the digital society is important to ensure the fairness of procedures through which goods or services are allocated[1] especially to marginalized groups and in order to enable them to voice their concerns and empower them to participate more fully in wider society.[2]

Digital inclusion can be defined as "effective participation of individuals and communities in all aspects of knowledge-oriented society and economic-oriented society by providing access to technology, by excluding availability barriers and enhancing the capability of communities to take ICT benefits"[3]. Scholars have identified several factors that contribute to inclusion, including access, digital literacy and digital skills[4].

Access is generally understood to be the precondition for participation in the digital society, even though it is by no means enough.[5] Identity management (i.e. a framework of policies and technologies to ensure that the right users have the appropriate access to technology resources) is an essential tool in ensuring 1) the participation of parents in distributed regulation of the platform, and 2) the availability of appropriate access for each user (e.g. teachers, students, parents, inspectors) according to their role. In the Shaad platform only students were recognized as users and in many cases siblings of different ages used the same device. For example, there are reports that older siblings used the platform to contact students from other schools (and often the opposite sex) because of the absence of a mechanism to ensure the logging out from their siblings' account who used the same device to access the Shaad. Moreover, parents had no means of ensuring appropriate use of the platform and occasionally had to use it to contact other parents or teachers as well. This also weakened the regulatory process because it could not be established who had used the device if an inappropriate message was sent. An SMC could play an effective arbitration role (in cases of misconduct) and push further development of the platform towards better identity management and inclusion of parents.

In addition, parents, teachers and other third parties should be aware of the capabilities and features of the platform, for example the type of content that can be uploaded and different options that one can use to communicate with other users. Otherwise, not everyone will be in a position to have an equal role in

---

[1] e.g. Azmi, A., Ang, Y. D., & Talib, S. A. (2016). Trust and justice in the adoption of a welfare e-payment system. *Transforming Government: People, Process and Policy*, *10*(3), 391-410.

Martin, A., & Taylor, L. (2021). Exclusion and inclusion in identification: Regulation, displacement and data justice. *Information Technology for Development*, *27*(1), 50-66.

Masiero, S., & Arvidsson, V. (2021). Degenerative outcomes of digital identity platforms for development. *Information Systems Journal*, *31*(6), 903-928.

[2] Maier, S., & Nair-Reichert, U. (2007). Empowering women through ICT-based business initiatives: An overview of best practices in e-commerce/e-retailing projects. *Information Technologies & International Development*, 4(2), 43-60;

Hassanin, L. (2008, September). Egyptian women artisans: ICTs are not the entry to modern markets. In *IFIP International Conference on Human Choice and Computers* (pp. 179-190). Springer, Boston, MA.

[3] European Commission. (2007). *European i2010 initiative on e-Inclusion: "To be part of the information society"*. Brussels: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM(2007) 694 final.

[4] Madon, S., Reinhard, N., Roode, D., & Walsham, G. (2009). Digital inclusion projects in developing countries: Processes of institutionalization. *Information Technology for Development*, *15*(2), 95-107.

Van Dijk, J.A.G.M. (2005). The deepening divide: Inequality in the information society, Sage Publications, Inc

Heeks, R. (2022). Digital inequality beyond the digital divide: conceptualizing adverse digital incorporation in the global South. *Information Technology for Development*, *28*(4), 688-704.

[5] Armenta, Á., Serrano, A., Cabrera, M., & Conte, R. (2012). The new digital divide: the confluence of broadband penetration, sustainable development, technology adoption and community participation. *Information Technology for Development*, 18(4), 345-353.

Friederici, N., Ojanperä, S., & Graham, M. (2017). The Impact of Connectivity in Africa: Grand Visions and the Mirage of Inclusive Digital Development, *Electronic Journal of Information Systems in Developing Countries*, 79(2), 1-20.

the decisions made. Without a diversity-oriented approach towards SMCs, the marginalized or powerless groups for example those from minority ethnic groups will find little or no opportunity to develop skills, exercise their rights to preserve their values and culture or play an autonomous role in content moderation on the platform. If not properly configured, SMCs may conceal actual power structures[6] (e.g. certain powerful ethnic groups or individuals who own or control the platform can influence and bias the content presented on the platform) and fail to achieve their objective of bringing democratic values to the governance of social networks. On the other hand, inclusion of marginal groups can enhance their sense of ownership and belonging to a community, inside the platform as well as outside. For instance, students and parents from a minority group may contribute actively to the platform by sharing their work, insights and this can motivate them to continue engaging with the platform and contribute more. This feeling of responsibility in the platform's success and taking active steps to maintain the platform's quality will contribute positively to the platform's growth and development.

Furthermore, in order to ensure that all stakeholders are aware of the opportunities for involvement in the governance of social media platforms, SMCs must meet the highest standards of transparency regarding their own operations. This could be exemplified by the encouragement of content production and participation by marginal groups, to which we now tend.

SMCs Must Go Further to Promote Inclusion through Encouraging Content Production by Marginal Groups

If the content and information on the platform are produced and used only in ways decided by dominant groups, marginal groups will not be adequately represented on these platforms and very likely will play no role in SMCs. Therefore, in educational contexts, SMCs must not limit their role to enforcing guidelines. Rather, they should actively encourage the creation and curation of content produced by marginal groups in order to ensure that they remain visible and can have a voice when critical decisions are to be made. This also ensures the display of diversity of values which can help maintain an open mind in discussing specific cases.

Ideally, Different Roles Defined in the Educational Context Must be Represented in the SMC.

The composition of members of SMCs in educational contexts must represent the various roles defined in this context. Therefore, it is necessary that teachers, parents and officials have representation and can play an active role in the functioning of such a council. Whether or not these roles will be equal cannot be determined without a careful consideration of broader legal context related to educational platform and distribution of accountability. Students can also be represented in these councils, although they are unlikely to have any voting rights because they are not mature enough. Additionally, individuals with experience in educational technology, and experts in software development, data security, privacy and other technical fields who can provide insight and advice to the council in platform design, features and best practices can also play an active role in SMCs. Additionally, leaders from local or national educational organizations, advocacy groups and associations who can provide guidance on policy and legal issues related to the educational platform should be represented in SMCs.

It is only by adhering to these principles that SMCs may contribute to the promotion of democratic values in the educational context as well as in society at large.

---

[6] Kettemann M.C. (ed.). (2022). How Platforms Respond to Human Rights Conflicts Online. Best Practices in Weighing Rights and Obligations in Hybrid Online Orders, Hamburg: Verlag Hans-Bredow-Institut.

Social Media Councils and Gender (In)Equality: An Analysis of Decisions by Meta's Oversight Board

Lucinda Nelson
QUEENSLAND UNIVERSITY OF TECHNOLOGY, BRISBANE, AUSTRALIA

Abstract

This short paper examines the potential role of social media councils (SMCs) in addressing gender inequality through an analysis of three relevant decisions of Meta's Oversight Board, as a current example of a platform-specific SMC. This analysis reveals the important roles the Oversight Board plays in publicly exposing gender inequality and improving transparency in Meta's content moderation, and the features that allow the Oversight Board to perform these roles, being: opportunity for public comment; sufficient independence from Meta; the ability to ask Meta questions about its policies and practices; publicly available reasoning; access to independent research; and an external set of rules to guide decision-making. This case study also highlights the Oversight Board's inability to directly effect changes to platform policies and enforcement practices, and the challenges this presents when Meta's policies and practices are discriminatory. Finally, this paper calls for further research into: whether the Oversight Board, or a different model of SMC, should be empowered to make binding decisions about platform policies and enforcement practices; and the potential for SMCs to address other axes of inequality, such as race and sexuality.

Context

Social media platforms have become an important space for public expression, discussion and deliberation. To date, platforms have mostly taken an undemocratic approach to content moderation, making decisions that significantly impact users' expression without public consultation or explanation. Women and gender diverse people are disproportionately impacted by content moderation policies and enforcement practices. Their ability to express themselves online is limited directly by discriminatory policies and practices,[1] and by a lack of moderation of harmful content that seeks to silence them.[2] In this context, SMCs have emerged as a potential mechanism for the development of better approaches to content moderation, including in relation to gender inequality. Meta's Oversight Board is just one example of a SMC, but it reveals interesting insights which can be applied in different contexts and in the development of new SMCs.

Case study 1: Breast cancer symptoms and nudity[3]

---

[1] Salty. 2021. Algorithmic Bias Report. https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/.

[2] Khan, Irene. 2021. Promotion and Protection of the Right to Freedom of Opinion and Expression (No A/76/258, United Nations). https://documents-ddsny.un.org/doc/UNDOC/GEN/N21/212/16/PDF/N2121216.pdf?OpenElement.

[3] Oversight Board. 2020. 'Breast cancer symptoms and nudity.' https://www.oversightboard.com/decision/IG-7THR3SI1/.

This case involved the removal of an Instagram post containing photos of bare breasts with breast cancer symptoms, under the *Adult Nudity and Sexual Activity policy*.[4] Uncovered female nipples are generally prohibited under the *Adult Nudity and Sexual Activity policy* but are permitted if posted for 'educational or medical purposes', including breast cancer education. After this case was selected by the Oversight Board, Meta (at the time, Facebook) acknowledged the removal was a mistake, restored the post and urged the Oversight Board to decline the case on the basis that it was now moot. The Oversight Board rejected this argument, heard the case, and overturned the original decision to remove the post.

This case highlights the value of the Oversight Board in publicly exposing gender inequality in Meta's policies and enforcement practices. Content moderation is largely 'black box'.[5] Women and gender diverse people report being disproportionately targeted by content moderation removals,[6] but it is difficult to prove these claims without evidence of social media platforms' internal policies and practices.[7] By proceeding with the case, even after Meta had acknowledged the error and restored the post, the Oversight Board drew **public attention to gender inequalities** in automated content moderation. It found that Meta's reliance on automated enforcement of the *Adult Nudity and Sexual Activity policy* is likely to have a disproportionate impact on women, because of the different treatment of 'male' and 'female' nipples. The case also provided an **opportunity for public comments** on the topic, which included arguments that Meta's nudity policies discriminate against women. Importantly, this decision could only be made because the Oversight Board was **sufficiently independent** from Meta to reject the proposal to decline the case.

Case study 2: India sexual harassment video[8]

This case involved the removal of a video posted on Instagram depicting a Dalit woman in India being sexually assaulted, under Meta's *Adult Sexual Exploitation policy*.[9] The removal was flagged for review internally after an employee learnt about it on Instagram. Meta then restored the post, with a warning screen, under a newsworthiness allowance. The newsworthiness allowance is broad and rarely applied, and involves balancing 'the public interest' and the potential for harm, without clear criteria. The Oversight Board upheld Meta's eventual decision to restore the post to Instagram. This decision was selected for analysis as sexual assault disproportionately affects women and is a key site of gender inequality.

This decision reflects the value of the Oversight Board as a body for making complex, policy-shaping determinations more transparently. Traditionally, large social media platforms have taken a formal equality approach to content moderation, in part due to the difficulty of considering context at scale.[10]

---

[4] Meta. 'Adult nudity and sexual activity.' Accessed February 23, 2023. https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual-activity/.

[5] Gillespie, Tarleton. 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. New Haven: Yale University Press.

[6] West, Sarah Myers. 2018. 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms.' New Media & Society 20(11): 4366-4384; Salty. 2021. Algorithmic Bias Report. https://saltyworld.net/product/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/.

[7] Cotter, Kelley. 2021. '"Shadowbanning Is Not a Thing": Black Box Gaslighting and the Power to Independently Know and Credibly Critique Algorithms' Information, Communication & Society: 1-18.

[8] Oversight Board. 2020. 'India sexual harassment video.' https://www.oversightboard.com/decision/IG-KFLY3526/.

[9] Meta. 'Adult sexual exploitation.' Accessed February 23, 2023. https://transparency.fb.com/en-gb/policies/community-standards/sexual-exploitation-adults/.

[10] Bartolo, Louisa. 2021. '"Eyes Wide Open to the Context of Content": Reimagining the Hate Speech Policies of Social Media Platforms through a Substantive Equality Lens.' Renewal: A Journal of Social Democracy 29(2): 39-51.

In this case, the Oversight Board was able to take a more substantive equality approach, considering a range of contextual factors relevant to both the public interest value of the content and its potential to cause harm, including the particular marginalisation of Dalit women in India. This contextual understanding was supported by **independent research** commissioned for this case. Importantly for **transparency**, the case was opened for public comment, and the Oversight Board's **reasoning is publicly available** for people to consider and critique in a way that decisions made solely by Meta are not. While transparency alone does not equal accountability, it is a prerequisite.[11]

Where the existing rules and exceptions are not clearly defined, as in this case, the Oversight Board plays a significant role in shaping platform rules through its precedential decisions. In these cases, the composition of the Oversight Board could have a critical impact on the decisions made, as different people are likely to have different perspectives on the public interest value of a piece of content. Members holding explicit or implicit sexist views, for example, would be likely to interpret any ambiguities in platform policies in sexist ways. The Oversight Board's reference to international human rights law, as a pre-determined, **external set of rules**, is useful to mitigate the impact of member perspectives. For the purposes of advancing gender equality, it is also important that board members understand and are **genuinely committed to addressing gender inequality.**

Case study 3: Gender identity and nudity[12]

These two cases (bundled) involved the removal of two Instagram posts consisting of photos of a transgender person and a non-binary person, both bare chested, under the *Sexual Solicitation* community standard.[13] The Oversight Board overturned Meta's decisions to remove the posts in question on the basis that the nipples were not 'female' and were therefore permitted under an exception.

This decision demonstrates the role of the Oversight Board in improving transparency by **exposing inconsistencies** between Meta's internal guidelines and its public-facing policies. Using its unique position to **ask questions** of the company, the Oversight Board established that the internal reviewer guidance on sexual solicitation differed from Meta's written policy in a way that resulted in the incorrect removal of permitted content and had disproportionate negative impacts on women and gender diverse users.

This decision also highlights the notable limitations of the Oversight Board in effecting policy change. Numerous public comments and the Oversight Board itself raised serious concerns about Meta's *Adult Nudity and Sexual Activity* and *Sexual Solicitation* policies. These concerns included the confusing combination of multiple prohibitions and exceptions in the relevant policies, and the distinction between 'male' and 'female' nipples which disproportionately limits women's expression and presumptively sexualises 'female' nipples. The Oversight Board specifically found that Meta's policies create 'greater barriers to expression for women, trans and gender non-binary people.' However, while the Oversight Board can make recommendations to Meta, it **does not have authority to make binding decisions about policy changes.** As the Oversight Board is a decision-reviewing body, this may be appropriate from a democratic, separation of powers perspective, but leaves the issue of discriminatory policies

---

[11] Suzor, Nicolas, Sarah Myers West, Andrew Quodling and Jillian York. 2019. 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation', International Journal of Communication 13: 1526-1543.

[12] Oversight Board. 2022. 'Gender identity and nudity. https://www.oversightboard.com/decision/BUN-IH313ZHJ/. 'Sexual solicitation.' Accessed February 23, 2023. https://transparency.fb.com/en-gb/policies/community-standards/sexual-solicitation/.

[13] Meta. 'Sexual solicitation.' Accessed February 23, 2023. https://transparency.fb.com/en-gb/policies/community-standards/sexual-solicitation/.

unaddressed. **Further research** is therefore needed into how platforms should develop their policies to promote gender equality, including whether a **different model of SMC** may be appropriate for this purpose. This is a particularly difficult challenge in relation to content, including female nudity, that is acceptable in some cultures, contexts and locations, but not others.

Conclusion

This case study has revealed three main benefits of the Oversight Board in addressing gender inequality: (1) exposure of gender inequality in content moderation; (2) greater transparency around complex decisions; and (3) greater public involvement in decisionmaking.

These benefits have been facilitated by the following features of the Oversight Board, which should be adopted in the establishment of future SMCs: opportunity for public comment; sufficient independence from Meta; ability to ask Meta questions about its policies and practices; publicly available reasoning; access to independent research; and an external set of rules to guide decision-making. The Oversight Board has recently named gender as one of its strategic priorities,[14] and any relevant future decisions should be analysed for further insights.

Although this study has focused on Meta's Oversight Board, the main benefits and features identified are not platform specific. Provided they have adequate funds, any platform that performs content moderation could implement a SMC with similar features and would likely see similar benefits. These findings could also be applied to other SMC models, including multi-platform and multi-stakeholder SMCs.

Similarly, although this paper has focused on the issue of gender inequality, the benefits and features of the Oversight Board are not gender-specific and may be applicable to other types of structural inequality. Analyses of decisions relating to other types of inequality should be undertaken to investigate the potential of SMCs in these contexts.

This case study also highlighted a significant limitation of the Oversight Board: as a decision-reviewing body, it cannot directly address gender inequality in content moderation policies and enforcement practices. Further research is therefore needed into whether the Oversight Board, or a different model of SMC, should be empowered to make binding decisions about platform policies and practices.

---

[14] Oversight Board. 2022. 'Oversight Board announces seven strategic priorities.' https://www.oversightboard.com/news/543066014298093-oversight-board-announces-seven-strategic-priorities/.

Executive-appointed social media councils: A case study from India

Prateek Waghre
INTERNET FREEDOM FOUNDATION, NEW DELHI, INDIA

Abstract

This short paper presents a case study of the executive-appointed social media councils in India. First, the paper presents the context surrounding the subordinate legislation that enabled the creation of these councils. It proceeds to critique these councils based on legitimacy, lack of specificity and capacity. The paper hypothesises that in their current form, the councils may counterbalance platform power but, in the process, concentrate power in the hands of 'old school speech regulation' bodies.

Background

This section provides the context in which the rules that enabled the formation of the executive-appointed councils were notified. It highlights the tensions between certain social media platforms and the executive branch of the government, as well as the efforts of the latter to exert control over the internet.

In February 2021, two ministries of the Government of India (GoI), the Ministry of Electronics and Information Technology (MeitY) and the Ministry of Information and Broadcasting (MIB), held a joint press conference announcing the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (IT Rules, 2021).[1] The notified version of the IT Rules, 2021 would go into effect on 25th May, 2021, and were significantly different from the draft Information Technology [Intermediaries Guidelines (Amendment) Rules] 2018 (Intermediary Guidelines, 2018) that were open for public feedback in December, 2018. The Intermediary Guidelines were made available for consultation after press reports of close-door meetings and a draft version being published by civil society organisations.[2]

The IT Rules 2021 divided administration between MeitY and MIB. MeitY would administer Part II of the rules, pertaining to 'Due Diligence by Intermediaries and Grievance Redressal Mechanism(s)'. Part II of the rules included obligations that intermediaries would have to fulfil; the mechanism for a grievance redressal process which included requirements to appoint a Grievance Redressal Officer, timelines for acknowledgement and disposal of grievances. Part II also defined a new category of intermediaries, called significant social media intermediaries if the number of registered users in India

---

[1] IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 English, 25 February 2021, https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf. Press Information Bureau, India, 'Government Notifies Information Technology (Intermediary Guidelnes and Digital Media Ethics Code) Rules 2021', 25 February 2021, https://pib.gov.in/PressReleseDetailm.aspx?PRID=1700749.

[2] Seema Chisti, 'Govt Moves to Access and Trace All "Unlawful" Content Online', The Indian Express (blog), 24 December 2018, https://indianexpress.com/article/india/it-act-amendments-data-privacy-freedom-of-speech-fb-twitter-5506572/. 'India Must resist the lure of the Chinese model of online surveillance and censorship #IntermediaryRules #RightToMeme #SaveOurPrivacy, Internet Freedom Foundation, 24 December 2019, https://internetfreedom.in/india-must-resist-the-lure-of-the-chinese-model-of-surveillance-and-censorship-intermediaryrules-righttomeme-saveourprivacy/.

were higher than 5 million, as defined via gazette notification in February 2021. Significant social media intermediaries were also required to have in-country Chief Compliance Officers, Grievance Redressal Officers, and Nodal Officers. They were also obliged to enable 'traceability' of messages to the first originator of a message in India.

Part III of the IT Rules, 2021 were to be administered by MIB and were applicable to publishers of 'news and current affairs content', and 'online curated content'. The Rules proposed a three-tier grievance redressal mechanism, with the publisher forming the first level. The second level would consist of self regulating bodies, and the third, an oversight mechanism consisting of an Inter-Departmental Committee with powers to issue guidelines, advisories, order and directions to publishers.

Between January and June 2021, MeitY and Twitter were interlocked in a 'jawboning' exercise. In late January, Twitter had complied with and then reversed course on some content takedown orders issued by MeitY related to the hashtag 'ModiPlanningFarmerGenocide' stating that it would not take actions against accounts belonging to 'news media entities, journalists, activists, and politicians' as it would violate their 'fundamental right to free expression under Indian law'.[3] In early February, the Government of India threatened Twitter with penal action for not complying with content takedown orders.[4] An unnamed government source would claim that Twitter took down '90-95%' of the accounts.[5] Various government officials and ministries also created accounts and advertised their presence on Koo, an India-based microblogging service.[6] In May 2021, Twitter flagged a post by a spokesperson of the Bhartiya Janta Party (BJP) as containing 'manipulated media'.[7] MeitY wrote to Twitter expressing its objection.[8] The Delhi Police issued notice to Twitter India's Managing Director, Manish Maheshwari.[9] This was followed by a special cell of the Delhi Policy arriving at Twitter's office in New Delhi seeking information.[10] In June 2021, with the IT Rules, 2021 going into effect, Twitter's India MD was named in First Information Reports in at least 2 instances.[11]

---

[3] Twitter Safety. 'Updates on Our Response to Blocking Orders from the Indian Government'. 10 Febrzary 2021. https://blog.twitter.com/en_in/topics/company/2020/twitters-response-indian-government; Billy Perigo, 'Twitter Blocks Accounts Linked to India Farmers Protests | Time', 1 February 2021, https://time.com/5935003/india-farmers-protests-twitter/.

[4] Yuthika Bhargava. 'Farmers' Protests | Govt Issues Notice to Twitter on "Farmer Genocide" Hashtag. The Hindu. 3 February 2021. Sec. India, https://www.thehindu.com/news/national/farmers-protest-govt-issues-notice-to-twitter-warns-of-penal-action/article33739720.ece.

[5] Yuvrai Malik. 'Twitter Takes down "90-95%" Accounts in Line with Meity Orders: Govt Source | Business Standard News'. 12 February 2021, https://www.business-standard.com/article/current-affairs/twitter-takes-down-90-95-accounts-in-line-with-meity-orders-govt-source-121021200134_1.html.

[6] Niharika Sharma, 'The Indian Government Is Backing a Homegrown Alternative Because Twitter Won't Bend to Its Will', Quartz, 10 February 2021, https://qz.com/india/1970534/piyush-goyal-other-indian-ministers-promote-koo-app-on-twitter/.

[7] Special Correspondent, 'Twitter Flags Sambit Patra's Tweet on Congress "Toolkit" as Manipulated Media', *The Hindu*, 21 May 2021, sec. India, https://www.thehindu.com/news/national/twitter-flags-sambit-patras-tweet-on-congress-toolkit-as-manipulated-media/article34611486.ece.

[8] Yuthika Bhargava, 'Government Asks Twitter to Remove "Manipulated Media" Tag Fom Tweets Related to "Congress Toolkit"', The Hindu, 21 May 2021, sec. India, https://www.thehindu.com/news/national/government-asks-twitter-to-remove-manipulated-media-tag-fom-tweets-related-to-congress-toolkit/article34615696.ece.

[9] Live Law [@LiveLawIndia], 'Delhi Police on 21st May Issued Notice to Twitter MD Requesting His Presence on 22nd May in Connection with the "Congress Tool Kit" Matter @TwitterIndia @DelhiPolice #ToolkitCase @INClndia Https://T.Co/RAfHGsRZHW', Tweet, Twitter, 24 May 2021, https://twitter.com/LiveLawIndia/status/1396845158805630987.

[10] 'Police at Twitter's Door after BJP Posts Flagged; Opposition Slams "Intimidation"', The Indian Express (blog), 25 May 2021, https://indianexpress.com/article/india/twitter-india-delhi-police-raid-7328607/.

[11] Ismat Ara, 'Late Night FIR Against Twitter, Opposition Leaders, Journalists for Posts on Ghaziabad Attack', The Wire, 16 June 2021, https://thewire.in/government/late-night-fir-against-twitter-opposition-leaders-journalists-for-posts-on-ghaziabad-attack; Indu Bhan,'Delhi Police Receives Complaint against Swara Bhaskar, Twitter India MD in Ghaziabad Assault Case', *The Economic Times*, 17 June 2021, https://economictimes.indiatimes.com/news/india/delhi-police-receives-complaint-against-swara-bhaskar-twitter-india-md-in-ghaziabad-assault-case/articleshow/83597364.cms.

The IT Rules 2021 were criticised by civil society organisations on the basis of lack of constitutionality, exceeding the scope of the parent Information Technology Act, 2000, variation from the Intermediary Guidelines, 2018 and absence of subsequent public consultation, among others.[12] As of May, 2022, there were at least 17 challenges in various High Courts across India.[13] While the Supreme Court of India stayed proceedings in these cases, any interim orders were to remain in effect.[14] Certain clauses of Part III of the rules have been stayed by the High Courts of Kerala, Bombay and Madras. However, information revealed in response to Right to Information requests revealed that over 2000 news publishers had furnished details to MIB, even though the 3-tier mechanism had been stayed by the Bombay High Court. Facebook and Whatsapp have also challenged the traceability requirements before the Delhi High Court.[15]

Executive-appointed social media councils

In spite of the various challenges to the IT Rules, 2022, on 3rd June 2022, MeitY proposed, withdrew a set of amendments to the IT Rules 2021.[16] On 6th June 2022, it once again published the proposed amendments for public feedback prefaced by a Press Note which stated that 'early stage or growth stage Indian companies or Startups' would not be impacted, without specifying how.[17] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2022 (IT Amendment Rules, 2022) were officially notified in October 2022.[18] The amendments proposed the creation of one or multiple Grievance Appellate Committees (GACs) where users could appeal against decisions taken by the Grievance Officer of any intermediary, with decision of the GACs being binding. A GAC will consist of one chairperson and 2 full time members, and one of these members will be a member ex-officio. A GAC shall 'endeavour to resolve the appeal finally within thirty calendar days'. In June 2022, the Minister of State of Electronics and Information Technology (MoS-EIT) stated that despite appointing grievance officers intermediaries were not providing 'real redressal', which needed

---

[12] Archana Sivasubramanian and Manish, 'Unpacking the IT Rules, 2021', CPR (blog), 21 December 2021, https://cprindia.org/unpacking-the-it-rules-2021/; 'ANALYSIS OF THE INFORMATION TECHNOLOGY (INTERMEDIARY GUIDELINES AND DIGITAL MEDIA ETHICS CODE) RULES, 2021', SFLC.in, 27 February 2021, https://sflc.in/analysis-information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021/; Neeti Biyani and Amrita Choudhury, 'Internet Impact Brief: 2021 Indian Intermediary Guidelines and the Internet Experience in India', *Internet Society* (blog), accessed 12 February 2023, https://www.internetsociety.org/resources/2021/internet-impact-brief-2021-indian-intermediary-guidelines-and-the-internet-experience-in-india/; 'How the Intermediaries Rules Are Anti-Democratic and Unconstitutional.', Internet Freedom Foundation, 27 February 2021, https://internetfreedom.in/intermediaries-rules-2021/.

[13] 'Table Summarizing Challenges to IT Rules, 2021', Google Docs, accessed 12 February 2023, https://docs.google.com/document/d/1kmq-AIRO1XpPaThvesl5xQq2nVkZv6UdmaKFAJ8AMTk/edit?usp=embed_facebook.

[14] 'Supreme Court Stays Proceedings before High Courts Challenging IT Rules, 2021, Interim Orders to Continue', Internet Freedom Foundation, 9 May 2022, https://internetfreedom.in/supreme-court-stays-proceedings-before-high-courts-challenging-it-rules-2021-interim-orders-to-continue/.

[15] PTI, 'WhatsApp Challenges New IT Rules in Delhi HC, Terms It "Unconstitutional"', 26 May 2021, https://theprint.in/india/whatsapp-challenges-new-it-rules-in-delhi-hc-terms-it-unconstitutional/666023/.

[16] 'MeitY Abruptly Withdraws Proposed Amendments to Technology and Social Media Rules - ET Government', ETGovernment.com, 3 June 2022, https://government.economictimes.indiatimes.com/news/governance/meity-abruptly-withdraws-proposed-amendments-to-technology-and-social-media-rules/91974760; 'MeitY Publishes and Then Withdraws a Proposal to Amend IT Rules, 2021', Internet Freedom Foundation, 3 June 2022, https://internetfreedom.in/meity-publishes-and-then-withdraws-a-proposal-to-amend-it-rules-2021/.

[17] Ministry of Electronics and Information Technology, 'Press Note Dated 6 June 22 and Proposed Draft Amendment to IT Rules 2021', 6 June 2022, https://www.meity.gov.in/writereaddata/files/Press%20Note%20dated%206%20June%2022%20and%20Proposed%20draft%20amendment%20to%20IT%20Rules%202021.pdf.

[18] Ministry of Electronics and Information Technology, 'Notification Dated, the 28th October, 2022 G.S.R. 794(E): The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2022 | Ministry of Electronics and Information Technology, Government of India', 28 October 2022, https://www.meity.gov.in/content/notification-dated-28th-october-2022-gsr-794e-information-technology-intermediary-guidelines.

to be addressed.[19] Repeating this position in October 2022, the MoS-EIT added that the IT Amendment Rules, 2022 were meant to make the internet safer and that the government did not want to be ombudsmen for the internet and was doing so reluctantly.[20]

The Asia Internet Coalition proposed self-regulatory mechanisms instead of the GAC-model.[21] According to media reports, U.S.-India Business Council (USIBC), part of the U.S. Chamber of Commerce, and U.S.-India Strategic Partnership Forum (USISPF) internally discussed concerns such as independence of the GACs, lack of checks and balances, absence of civil society representation.[22] Civil society organisations and experts raised concerns about executive influence on GACs and their binding decisions about content, their suspect legality on account of being outside the IT Act, 2000, with some even calling for their withdrawal.[23] In January 2022, MeitY invited applications for full time membership of the GACs with a deadline of 12th January, 2022.[24] On 27th January, 2022 it notified the formation of 3 3-member GACs with each being chaired by a member of the Ministries of Home Affairs, Information and Broadcasting, Electronics and Information Technology, respectively.[25] An accompanying press release stated that committees will be functional from 1st March, 2023 after which users would be able to "appeal against (the) decision of the grievance officer of the social media intermediaries and other online intermediaries."[26]

Discussion and recommendations

---

[19] PTI, 'Social Media Platforms Not Adequately Redressing Grievances: Rajeev Chandrasekhar', Business Today, 23 June 2022, https://www.businesstoday.in/latest/policy/story/social-media-platforms-not-adequately-redressing-grievances-rajeev-chandrasekhar-338953-2022-06-23.

[20] Yuthika Bhargava, 'Amended IT Rules Are Meant to Make Web Safer for All: Rajeev Chandrasekhar', The Hindu, 29 October 2022, sec. Interview, https://www.thehindu.com/opinion/interview/govts-intent-is-to-work-with-social-media-intermediaries-not-be-seen-as-adversarial-it-mos/article66070407.ece; Aihik Sur and Deepsekhar Choudhury, 'IT Rules Amendment: Govt Doesn't Want to Be Internet Ombudsman, Says Rajeev Chandrasekhar', Moneycontrol, 31 October 2022, https://www.moneycontrol.com/europe/?url=https://www.moneycontrol.com/news/business/startup/grievance-panel-for-intermediaries-will-be-a-traffic-signal-for-user-appeals-rajeev-chandrasekhar-9417991.html; Aditi Agrawal, "'Doing This Reluctantly", Redressal Mechanism "Broken": IT Minister Explains Need for Changed Rules', Newslaundry, 29 October 2022, https://www.newslaundry.com/2022/10/29/doing-this-reluctantly-redressal-mechanism-broken-it-minister-explains-need-for-changed-rules.

[21] Anushka Jain, 'IT Rules 2021: Submission on Safe Harbour Status, Enforcing Compliance, Other Issues', MediaNama (blog), 2 August 2022, https://www.medianama.com/2022/08/223-asia-internet-coalition-submission-on-it-rules-gac-compliance/.

[22] Aditya Kalra and Munsif Vengattil, 'U.S. Lobby Groups Cast Doubts over Independence of India Content Appeal Panel | Reuters', 20 July 2022, https://www.reuters.com/technology/us-lobby-groups-cast-doubts-over-independence-india-content-appeal-panel-2022-07-20/.

[23] Namrata Maheshwari Chima Raman Jit Singh, 'Civil Society Calls on Indian Government to Withdraw Amendments to IT Rules', Access Now (blog), 12 july 2022, https://www.accessnow.org/press-release/india-it-rules-amendments-joint-submission/; Aarathi Ganesan, 'Experts Flag Free Speech and Self-Censorship in India's Amended IT Rules', MediaNama (blog), 22 June 2022, https://www.medianama.com/2022/06/223-it-rules-amendments-india-free-speech-big-tech/; Tejasi Panjiar, 'A Public Brief on the IT Amendment Rules, 2022 a.k.a "How the Government Is Trying to Moderate Online Speech"', Internet Freedom Foundation, 10 November 2022, https://internetfreedom.in/public-brief-on-the-it-amendment-rules-2022/.

[24] Ministry of Electronics & IT [@GoI_MeitY], '#Hiring! Applications Are Invited for the Appointment of Full-Time Members of GAC(s). To Apply, Send an Email with a Scanned Copy of Your Duly Completed Application Form and CV to Group Coordinator (Cyber Law Division), @GoI_MeitY at Cyberlaw-Legal@meity.Gov.in. #DigitalIndia Https://T.Co/C6hp5cM3iH', Tweet, Twitter, 3 January 2023, https://twitter.com/GoI_MeitY/status/1610143208000081926; Digital India [@_DigitalIndia], '#Hiring! Applications Are Invited for the Appointment of Full-Time Members of GAC(s). To Apply, Send an Email with a Scanned Copy of Your Duly Completed Application Form and CV to Group Coordinator (Cyber Law Division), @GoI_MeitY at Cyberlaw-Legal@meity.Gov.in. #DigitalIndia Https://T.Co/R3RLwtLUx4', Tweet, Twitter, 3 January 2023, https://twitter.com/_DigitalIndia/status/1610143064076742659.

[25] Ministry of Electronics and Information Technology, 'Establishment of Grievance Appellate Committees under Rule 3A of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 | Ministry of Electronics and Information Technology, Government of India', accessed 13 February 2023, https://www.meity.gov.in/content/establishment-grievance-appellate-committees-under-rule-3a-information-technology.

[26] Press Information Bureau, India, 'Three Grievance Appellate Committees (GACs) Notified on the Recently Amended "IT Rules 2021"', 28 January 2023, https://pib.gov.in/pib.nic.in.

This section critiques the GACs on the basis of legitimacy, lack of specificity and capacity. It considers aspects such as the consultation process, legislative uncertainty, operational transparency, vagueness of remit as well as the ability, in terms of scale and expertise, to adjudicate the volume and likely complex nature of appeals expected through the process.

Legitimacy

The GACs draw their basis from a set of rules that had a significant impact on how people in India interact with services on the internet. Therefore the adherence to an open, transparent and responsive public consultation process should have been a crucial component of their drafting. However, the Intermediary Rules, 2018 were released for public consultation after media reports of closed-door meetings, and a version being released in the public domain. These rules did not explicitly seek the appointment of grievance redressal officers, nor oversight of 'news and current affairs content', and 'online curated content'. The IT Rules, 2021 were announced and went into effect in 3 months without any public consultation in their final form.

In March 2021, the Global Network Initiative (GNI) had written a letter to the then Union Minister for Electronics and Information Technology calling on MeitY to 'consider revising the rules and engage in an open, deliberative process about how to address and mitigate these concerns'.[27] Civil society organisations flagged that the process was in contravention of the Pre-legislative Consultation Policy, 2014, lacked a principled approach and were instead driven by political economy, and even called for their withdrawal.[28] Instead of engagement, the minister stated in an interview that they were based on prior consultations, committee reports and court rulings.[29]

Further, multiple civil society organisations also expressed the position that the IT Rules, 2021 were unconstitutional, and went beyond the scope of the parent act.[30] Since the IT Amendment Rules, 2022 are based on the IT Rules, 2021, this assessment extends to them. These concerns were reemphasised specifically with regard to the GACs when they were proposed and subsequently notified stating that there was legislative uncertainty as an executive-appointed council could decide on matters related to free speech based on grounds that were not stated under Section 69A of the IT Act, 2000 or Article 19(2) of the Constitution of India.[31]

Lack of specificity and capacity

---

[27] 'GNI Analysis: Information Technology Rules Put Rights at Risk in India', Global Network Initiative, accessed 14 February 2023, https://globalnetworkinitiative.org/india-it-rules-2021/.

[28] 'ANALYSIS OF THE INFORMATION TECHNOLOGY (INTERMEDIARY GUIDELINES AND DIGITAL MEDIA ETHICS CODE) RULES, 2021'; Archana Sivasubramanian and Manish, 'Unpacking the IT Rules, 2021'; 'Dear MEITY, Withdraw the New IT Rules!', Internet Freedom Foundation, 23 March 2021, https://internetfreedom.in/withdraw-the-it-rules/.

[29] Aashish Aryan, '"Our Commitment to Privacy Is Unimpeachable. Are They Permitting Free Speech by Not Obeying Constitution?": Ravi Shankar Prasad', *The Indian Express* (blog), 29 May 2021, https://indianexpress.com/article/business/economy/ravi-shankar-prasad-it-rules-privacy-twitter-whatsapp-7334837/.

[30] 'How the Intermediaries Rules Are Anti-Democratic and Unconstitutional.'; Torsha Sarkar et al., 'On the Legality and Constitutionality of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021', The Centre For Internet & Society, 21 June 2021, https://cis-india.org/internet-governance/legality-constitutionality-il-rules-digital-media-2021.

[31] Ganesan, 'Experts Flag Free Speech and Self-Censorship in India's Amended IT Rules'; Tejasi Panjiar, 'A Public Brief on the IT Amendment Rules, 2022 a.k.a "How the Government Is Trying to Moderate Online Speech"'.

Table 1 collates the approximate number of decisions taken by a specific set of significant social media intermediaries based on user reports and self reported numbers for proactive action or action taken. Between October and December 2022, Sharechat received nearly 8 million user reports.[32] In the same period Koo received nearly 30000 user reports and took more than 80,000 proactive content moderation decisions, Facebook and Instagram took action against over 65 million and nearly 10 million pieces of content respectively, and Snap Inc. reported receiving over 550,000 content and account reports.[33] If appeals against even 0.1% of these actions make it to the GACs, they would have to deal with tens of thousands appeals on a monthly basis. Meta's oversight board, with an operating budget of over 100 million dollars has received over 2 million appeals and picked up only 42 cases.[34]

Approximate number of decisions taken based on user reports and action taken disclosures

Social Media Platform	October 2022	November 2022	December 2022
Facebook	over 27 million	over 18 million	over 21 million
Instagram	over 2.6 million	over 3.1 million	over 3 million
Sharechat	over 3.9 million	over 1.8 million	over 2.1 million
Snap Inc.	over 195,000	over 180,000	over 175,000
Koo	over 29000	over 34000	over 16000

Table 1: Compiled by author based on IT Rules, 2021 compliance disclosures.

The press release announcing the constitution of 3 GACs stated that they would 'endeavour to resolve the appeal finally within thirty calendar days'. It does not explicitly state whether the GACs will adjudicate all the appeals or have discretion over which cases they choose. If so, there are immediate questions about their ability to handle the potential volume of cases. Alternatively, if the GACs are meant to have discretion over which cases to pick, then any guiding criteria for doing so have not been defined, thus also raising the risk of arbitrary or motivated case selection. The presence of ex-officio members, who are each currently chairpersons of their respective GACs raise questions about the independence of the GACs.

While the press release states that '(p)eriodic reviews of GACs and reporting and disclosures of GAC orders' will be a part of the process, there has been no further information about the specifics of what these reviews and disclosures will contain, nor the frequency of any such reports. In addition, no operating budgets have been specified. While the conversation mainly revolves around social media intermediaries, the GACs will also be the point of appeal for decisions taken by grievance redressal officers across all kinds of intermediaries. This vagueness of remit creates potential for the GACs to make content moderation decisions at different layers of the technology stack as well.

There is also no clarity on the basis of allocation of appeals across the 3 GACs and what implications that may have for the grievance filing process. Though the IT Amendment Rules, 2022 allow for the assistance from person(s) 'having requisite qualification, experience and expertise in the subject matter', the inadequate representation of civil society, academia, professionals with expertise in trust and safety

---

[32] 'ShareChat Transparency Reports', accessed 14 February 2023, https://help.sharechat.com/transparency-report/.

[33] 'Koo: View the Latest Koos » Monthly Compliance Reports', accessed 14 February 2023, https://info.kooapp.com/monthly-compliance-reports/; 'Regulatory and Other Transparency Reports | Transparency Center', accessed 14 February 2023, https://transparency.fb.com/data/regulatory-transparency-reports/; 'India Transparency & Data | Snapchat Transparency', accessed 14 February 2023, https://values.snap.com/en-GB/privacy/transparency/india.

[34] Steven Levy, 'Inside Meta's Oversight Board: 2 Years of Pushing Limits | WIRED', WIRED, 8 November 2022, https://www.wired.com/story/inside-metas-oversight-board-two-years-of-pushing-limits/; 'The Oversight Board | Transparency Center', accessed 14 February 2023, https://transparency.fb.com/en-gb/oversight/.

positions, etc. mean that GACs may not be adequately equipped to adjudicate what are likely a high volume of complex appeals.

Implications and recommendations

The proposed binding nature of the decisions combined with the general lack of transparency and specificity in a contested, polarised sphere such as social media leave the mechanism vulnerable to executive and ideological capture. The overall approach of the GACs also reflects an approach to content moderation that is neither suitable nor capable of scaling to meet the many challenges in today's information ecosystem. It relies on highly context-specific decisions taken about individual pieces of content, which may or may not have any precedent-setting value, to attempt to address systemic issues that are caused by broader societal-level problems. Aggregation of individual decisions may not be able to address underlying problems since they are neither repeatable nor broadly applicable, given the complexities involved.[35]

Based on the approach to consultative processes, the overreach of subordinate legislation with the IT Rules, 2021 and subsequent amendments; low levels of operational transparency; vagueness of remit across multiple bodies and kinds of intermediaries; as well as questions surrounding the committees in terms of ability, from the perspectives of scale and capability; strong and sustained adherence to principles like human rights, rule of law and democratic values is unlikely. For these reasons, in the Indian context, an executive-appointed and thereby executive-controlled council is an undesirable intervention.

Due to the expected volumes, context-specific nature of content reports and disputes on social media mean even councils that are multi stakeholder in nature can be expected to find it difficult to address problems at scale. Decision-making for councils is likely to be further complicated by the tendency of adversarial groups to employ techniques like malign creativity to evade detection, accountability and introduce a layer of plausible deniability.[36] For example, a combination of the names of two taxi aggregators in India are often employed as an anti-minority dog whistle.[37] There is a need to better understand the impact of social media-based communication on collective behaviour.[38] Thus even multi-stakeholder councils should be approached conservatively, and not result in diversion of resources for appropriate research into understanding their impact.

---

[35] Evelyn Douek, 'Content Moderation as Systems Thinking', *Harvard Law Review* 136, no. 2 (2022), https://harvardlawreview.org/2022/12/content-moderation-as-systems-thinking/.

[36] Nina Jankowicz et al., 'Malign Creativity: How Gender, Sex, and Lies Are Weaponized Against Women Online | Wilson Center', accessed 17 February 2023, https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online; Mohua Das, 'How Desi Troll Armies Have Built a Coded Language of Abuse | India News - Times of India', 8 February 2022, https://timesofindia.indiatimes.com/india/how-desi-troll-armies-have-built-a-coded-language-of-abuse/articleshow/89406114.cms; Aishwarya Varma, 'Can Tech and Humans Work Together To Make Social Media Less Communally Charged?', TheQuint, 27 April 2022, https://www.thequint.com/news/webqoof/communal-language-and-moderation-social-media-india.

[37] Prateek Waghre, prateekwaghre@mastodon.social [@prateekwaghre], 'It Amazes Me That the Names of the Two Leading App-Based Cab Aggregators in India Have Been Combined in Word Play That Serves a Bigoted End. Malign Creativity" Indeed'. Tweet, *Twitter*, 18 August 2021, https://mobile.twitter.com/prateekwaghre/status/1428011618105913347.

[38] Joseph B. Bak-Coleman et al., 'Stewardship of Global Collective Behavior', *Proceedings of the National Academy of Sciences* 118, no. 27 (6 July 2021): e2025764118, https://doi.org/10.1073/pnas.2025764118.
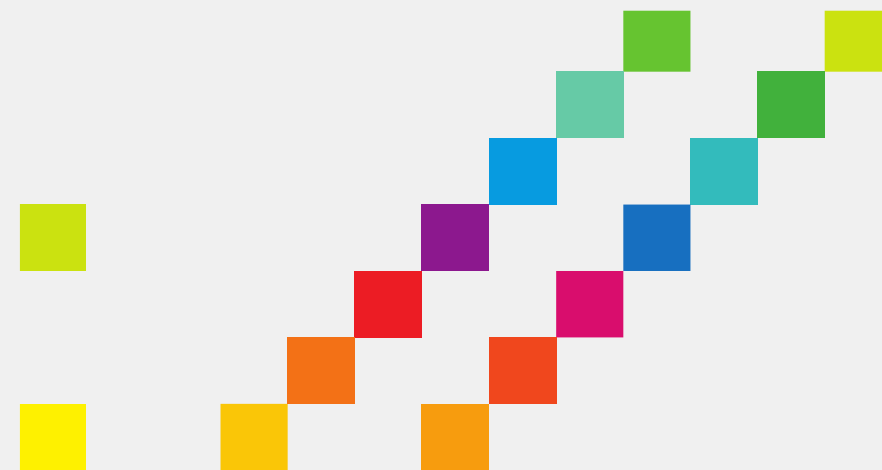
MATTHIAS C. KETTEMANN, JOSEFA FRANCKE, CHRISTINA DINAR AND LENA HINRICHS (EDS.)

Platform://Democracy

Research Report Europe

Platform://Democracy

Perspectives on Platform Power, Public Values and the Potential of Social Media Councils: Research Report Europe

edited by Matthias C. Kettemann, Josefa Francke, Christina Dinar and Lena Hinrichs

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT

ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY

STIFTUNG MERCATOR

Contributors

Name(s)	Affiliation
Pierre François Docquir	Independent Researcher
Rachel Griffin	SciencesPo
Bettina Hesse	Verdi
Niklas Eder	Meta Oversight Board
Dominik Piétron	Humboldt-Universität zu Berlin
Clara Iglesias Keller & Theresa Züger	Leibniz Institute for Media Research	Hans-Bredow-Institut & Alexander von Humboldt Institute for Internet and Society
Aleksandra Kuczerawy	KU Leuven
Pierre François Docquir	Independent Researcher
Rachel Griffin	SciencesPo

Table of Contents

Towards More Legitimacy in Rule-Making

Matthias C. Kettemann and Josefa Francke
LEIBNIZ INSTITUTE FOR MEDIA RESEARCH | HANS-BREDOW-INSTITUT, HAMBURG, GERMANY
ALEXANDER VON ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

The idea of creating platform councils to support the legitimacy of rules and recommender algorithms on social media platforms has gained traction in recent years. The concept involves establishing independent bodies, comprised of experts and stakeholders, to provide oversight and guidance to social media platforms on the development and implementation of their policies.

These councils can be responsible for developing, and ruling on, content moderation rules, data privacy guidelinces, and algorithmic transparency. They could also be tasked with conducting audits and reviews of platform policies and practices, and making recommendations for improvements.

One of the main goals of platform councils is to enhance the legitimacy and transparency of social media platforms' decision-making processes, and to increase public trust in their policies and practices. By involving independent experts and stakeholders in the development and implementation of platform policies, the hope is that social media platforms can create fairer rules and improve their algorithmic governance practices.

The following studies offer perspectives on the potentials of social media councils from European researchers.

Pierre François Docquir has utilized his research experience with ARTICLE 19 to emphasize the significant gap between local contexts and global company policies. He suggests that Social Media Councils (SMCs) fuelled by independent multi-stakeholder coalitions at the national level, may be able to bring local voices into content moderation, thereby bridging this gap.

Rachel Griffin takes a feminist perspective in her examination of SMCs, questioning how they can address existing online communication structures that perpetuate the marginalization of vulnerable groups. She emphasizes the importance of ensuring that SMCs' normative setup does not ignore structural problems, maximally includes diverse voices, and exercises sufficient power to demand substantive changes in business practices.

Bettina Hesse explores broadcasting councils as a potential role model for SMCs, identifying areas where they fall short and what SMCs can learn from them. Hesse highlights unchecked power relationships within the councils, with political parties and state representatives overrepresented at the expense of civil society. Additionally, the composition of most broadcasting councils lacks diversity and shows itself to be random in its choice of members. Hesse argues that broadcasting councils are quite removed from general society and should be more independent from overarching media organizations.

Niklas Eder's focus is on DSA regulation with regard to systemic risks. He suggests that platforms are required to do risk assessment and develop mitigation measures to mitigate the harm of platforms. However, given the conflict of interests, the platforms may not be the best-suited actors to do these assessments, and public institutions' involvement could conflict with their obligation to human rights such as freedom of expression. Eder argues that involving stakeholders and civil society via a Social Media Council could be the best way to evaluate risk and design effective mitigation measures.

Dominik Piétron argues for the need for collective data governance in the context of large-scale data extraction. He suggests that individual rights of informational self-determination are not sufficient for adequate protection. Piétron uses the example of a German city platform to demonstrate that SMCs

could be a suitable institution for such collective governance. Piétron proposes inclusive discussions and feedback opportunities, democratic accountability, algorithmic transparency, and effective implementation as design criteria.

Clara Iglesias Keller and Theresa Züger debate whether SMCs are suitable to combat democratic deficits in platform governance or merely validate existing power structures. They argue that citizens' roles and ways of contributing to democracy have changed under online conditions. Drawing on the concept of "digital citizenship," Keller and Züger suggest that SMCs should focus on improving input legitimacy, i.e., public participation.

Aleksandra Kuszerawy examines the potential role of SMCs within the Digital Services Act framework. She notes that internal complaint-handling mechanisms lack a crucial requirement for SMCs, namely independence. However, external out-of-court dispute settlement bodies could be congruent with the idea of independent SMCs, even if limited in their role to reviewing individual content moderation decisions. Kuszerawy maintains that independence and signaling of independence are necessary for legitimacy. Assessing the potential of the Meta Oversight Board to become an out-of-court dispute settlement body under the DSA, Kuszerawy concludes that the current OB structure does not allow for the necessary scaling up to meet the demands of the DSA.

Riku Neuvonen sheds light on the nature of SMCs as self-regulatory bodies and their relevance in the present debate around regulation of online communication. He draws on past instances where self-regulation has been deemed necessary and/or useful, particularly in the context of media councils.

In his analysis of SMCs Giovanni Di Gregorio approaches the issue through the lens of the European constitutional framework. He argues that, as long as SMCs are regarded as private entities, the application of constitutional principles is limited within their decision-making processes. Although SMCs have the potential to enhance oversight in the context of social media platforms, this is not a given, especially if the public has no involvement in the development of their procedures. Di Gregorio further notes that the global scale of SMCs may lead to a conflict between the adoption of international human rights law as a standard and local values and regional variations in its application.

Bringing local voices into conversations about content moderation

Pierre François Docquir
INDEPENDENT RESEARCHER, FORMER HEAD OF MEDIA FREEDOM PROGRAMME OF ARTICLE 19

In most countries of the world, looking at social media companies from the point of view of local actors is tantamount to contemplating distant planets that in any conceivable way appear to be out of reach. The same national observers, however, are acutely aware of the impact that online content has on the society they live in. In this contribution, I argue that, while content moderation processes often fail to integrate a robust understanding of the local circumstances in which online content deploys its effect, national stakeholders (such as for instance civil society organisations, media and academics) are well positioned to bring this specific knowledge and expertise to global companies, which would contribute to a more effective alignment of content moderation outcomes with the realities lived at the local or national level. I then look at two different possible approaches to enabling local voices to partake in conversations about content moderation and argue that an independent, multi-stakeholder and diverse Social Media Council (as defined in the model developed by free speech organisation ARTICLE 19) presents distinctive advantages over a consultative forum operated by a single company. Throughout the text, I'm using the terms local and national as almost interchangeable. Whether based on practical considerations (such as safety or the state of civil society), a 'local Social Media Council' could end up being set up as a national, sub-national or even regional mechanism is not the point of focus here: the goal is to shed light on the question of how to bridge the flagrant gap between global companies and many of the societies impacted by the massive circulation of online content.

The role of local voices in content moderation

While the content rules of social media platforms are elaborated as global norms, the resolution of any content moderation case inevitably requires an informed understanding of the circumstances in which a particular instance of speech has been uttered: in order to decide whether what is being said is in violation of, for instance, rules that prohibit speech that incites violence or discrimination, it is necessary to understand the language in which a particular message is expressed, as well as the social, cultural, historical and political dimensions of the societal context in which this message is disseminated. As research shows, a message that may seem innocuous when seen from Silicon Valley might contribute to reinforcing polarisation and a climate of violence in a complex and diverse country whose history has been defined by frequent episodes of violence. Content moderation systems that ignore local circumstances may cause severe societal harms (such as real-world violence, increased polarisation between social groups, or the undermining of trust in democratic institutions and electoral processes) or aggravate the risk of such societal harms (ARTICLE 19, 2022). In a normative perspective, the Santa Clara Principles on Transparency and Accountability in Content moderation (https://santaclaraprinciples.org/), a global civil society initiative aimed at encouraging global companies to comply with their responsibilities to respect human rights, set forth a principle of 'Cultural Competence' that requires that 'those making moderation and appeal decisions understand the language, culture, and political and social context of the posts they are moderating.' The principle entails that content moderation decisions need to be informed by sufficient understanding of the linguistic, cultural, social, economic and political dimensions of the relevant local or regional context.

However, it appears that content moderation systems – either human or algorithmic, or a combination of both – are often not designed to take local circumstances into consideration. More exactly, social media platforms seem to focus their attention and resources on a very limited number of countries (Newton, 2021). Other countries end up being entirely ignored: this means for instance that no locally-relevant classifier is built into the automated content moderation systems to detect problematic content (such as misinformation or hate speech) and that no partnership is established with local fact-checkers (TBIJ, 2022).

Local stakeholders could naturally play a role in helping social media companies integrate a solid understanding of the local context in their content moderation systems (Access Now, 2020). Partnerships, including regular meetings, with in-country organisations that have a deep knowledge of conflict dynamics could help identify and resolve cultural and social barriers to content reporting, and develop mitigating interventions in response to problematic content (SFCG, 2021). Multi-stakeholder coalitions at the national level could empower local communities to monitor and detect speech that incites violence or discrimination, and a regular cooperation between such coalitions and social media companies could contribute to ensuring that content moderation is aligned with the local context (UNESCO 2021). Regular and transparent consultations with a local coalition that would form a critical mass of local actors could provide social media companies with an effective approach to ensuring that their content moderation processes are better informed by a robust understanding of the local context. Both sides would benefit from such regular dialogue: in exchange for providing a valuable service to social media companies, national stakeholders would find themselves in a position where they can better achieve their own goals in relation to content moderation (ARTICLE 19, 2022a)

As has been documented by research, individuals and organisations who would be particularly well positioned to reflect on the meaning and potential impact of problematic speech disseminated in their society, often find themselves unable to engage with giant global entities that operate on a global scale. They are confronted with various obstacles that range from the absence of translation of the content rules, the lack of effectiveness of internal remedies provided by platforms, or the sheer inability to contact a representative of the company. Even civil society organisations that partake in partnerships with social media platforms (such as trusted flaggers' programmes) often report that they find it uneasy to make their viewpoints heard by their giant partner (ARTICLE 19, 2022a). And in this way, there remains a wide gap between global social media companies and local communities, while the moderation processes often remain out of touch with the specific societal context in which online content deploys its impact.

The participation of local voices in conversations on content moderation

As has been suggested, a form of dialogue between social media companies and local stakeholders could serve to better align content moderation with an informed understanding of the local or national context. To shed light on how this interaction could be designed, I will compare two different approaches: a consultative community forum recently announced by Meta and the model of Social Media Council that was developed by global free speech organisation ARTICLE 19 (2021). And I will argue that an independent Social Media Council (SMC) that represents the diversity of society would be in a better position to ensure an effective participation of local stakeholders than a consultative forum set up – and controlled – by social media companies.

Meta's Community Forum is described as a 'new method for making decisions' in the development of apps and technologies (Facebook, 2022). Drawing on the model of deliberative citizen assemblies, the Community Forum is meant to gather 'nearly 6,000 persons from 32 countries'. Organised by a global

social purpose company and a research centre of Stanford University, this global forum took place in December 2022 on a virtual closed platform that enables participants to discuss amongst themselves and to have access to experts. The non-binding recommendations of the forum will be made public at the end of the process.

The fact that the initiative is ongoing at the time of writing should not forbid initial thoughts on how it could contribute to solving the equation of integrating local context into content moderation processes. In this perspective, the forum certainly appears to be open to viewpoints from a broad diversity of countries. The partnership with expert organisations contributes to the legitimacy and credibility of the consultative process. However, for the specific matter discussed in this contribution, the community forum's main weakness lies in that it aims at making decisions globally, in a 'one size fits all' perspective. In addition, and even without looking at how participants are selected, its inclusivity inevitably remains limited: many countries, many languages, many cultures will simply not be represented in the deliberations. This being said, we could imagine a similar mechanism set at the national or regional level: assuming that participants would be selected in a manner that ensures a fair and transparent representation of the broad diversity of the living forces of society, such a deliberative process that combines learning (through access to expertise) with discussions could potentially deliver relevant recommendations on the alignment of content moderation with local circumstances. But there are two elements that could undermine the credibility and legitimacy of the process. First, the recommendations are not binding for the social media company, which ultimately remains free to base its decisions on whichever considerations it deems relevant. And the forum is commissioned, designed and controlled by a social media company: the participants do not have the opportunity to discuss the process itself. It also means, as illustrated by the recent disbandment of Twitter's Trust and Safety Council, that the company can at will bring an end to the forum, even very abruptly so (Former Members, 2022).

While a Social Media Council (SMC) could offer a comparable cocktail of deliberation and expertise (either through its own membership or external consultation), it presents significant differences. As defined in ARTICLE 19's approach (2021) an SMC is a multi-stakeholder mechanism that would oversee content moderation decisions on the basis of international human rights law both through the review of individual cases and the elaboration of general recommendations. While there have been debates on whether the SMC should be instituted as a global forum or as a national mechanism (GDPi, 2019). ARTICLE 19 argued in favour of a national SMC, except in situations where safety concerns would prevent the feasibility of such initiative.

For the purposes of this contribution, a first distinctive feature is that the SMC is meant to be created and operated by all the participating stakeholders, which means that its rules, processes and even its mere existence are not left to the discretion of a social media company alone – or, if it is thought of as a multi-platform council, of the participating social media companies. The funding of the SMC, which would come either from contributions of social companies, public funds or donations from international foundations, would need to be organised in a manner that prevents threats to its independence. A second point of contrast with a consultative Community Forum comes from the fact that the outcomes of the SMC's deliberations would be binding for the participating social media companies, which would have committed to executing its decisions in good faith. This ensures that stakeholders' voices could effectively carry some weight in terms of influencing content moderation practices. And finally, the ARTICLE 19 model proposed that the SMC should represent the whole diversity of society, including marginalised groups, which would confer additional democratic legitimacy to its deliberations.

While they present strong features, SMCs are not easy policy proposals. That is in part because legislators and stakeholders may be reluctant to the idea that the industry would take part in the definition of the rules that govern its behaviour, and that this very participation alongside other actors and under a regime of publicity, would ensure that compliance slowly percolates through the sector's practices. The complexity also comes from the need to resolve issues that are merely touched upon here, such as ensuring that funding does not undermine its sustainability and its independence, or the need to achieve an effective balance of powers between the various groups of stakeholders. The difficulties in setting it up notwithstanding, a national SMC could enable local stakeholders to bring their knowledge of the local circumstances to the analysis of content moderation issues affecting their society. In this model, the complexity of the local context would be analysed and discussed in the deliberations of a representative and diverse group of stakeholders, which ensures a higher degree of credibility and legitimacy than a consultative forum with limited representativeness. The same could indeed be said of the application of international human rights law to content moderation. In a multi-stakeholder Social Media Council, a social company would not be in a position to discreetly determine what the requirements of freedom of expression, privacy or due process are: instead, the definition of the rules that will govern content moderation would result from a collective deliberation of all the stakeholders that are impacted by the circulation of online content. The independence of the SMC and the (voluntarily) binding nature of its decisions reinforce its capacity to serve as an effective enabler for local stakeholders to contribute to the alignment of content moderation processes with the needs and realities of the society they live in.

At the time where a new European legislative framework for the digital world is entering into force, it should be noted that, as discussed elsewhere in this series, a national SMC could correspond to the definition of the out-of-court settlement mechanism provided for in Article 18 of the DSA. In any case, an effective and credible participation of local stakeholders in content moderation would provide a response to the disconnect between global companies and the societies that they impact: in the light of whistleblowers' revelations and of chaotic change of ownership, a local SMC could be key to restoring users' trust in social media platforms.

Sources

Access Now, https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf

ARTICLE 19, Content moderation and local stakeholders in Indonesia, June 2022, https://www.article19.org/wp-content/uploads/2022/06/Indonesia-country-report.pdf

ARTICLE 19, Bridging the Gap, June 2022a, https://www.article19.org/wp-content/uploads/2022/06/Summary-report-social-media-for-peace.pdf

ARTICLE 19, Social Media Councils: one piece in the puzzle of content moderation, Oct. 2021, https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf

Facebook, Improving People's Experiences Through Community Forums, Nov. 2022, https://about.fb.com/news/2022/11/improving-peoples-experiences-through-community-forums

Former members of the Twitter Trust and Safety Council, Statement, 14 Dec. 2022, https://www.article19.org/resources/twitter-disbanding-of-trust-and-safety-council

GDPi, From Concept to Reality, June 2019, https://cyber.fsi.stanford.edu/gdpi/content/social-media-councils-concept-reality-conference-report

Newton, C., The tier list: how Facebook decides which countries need protection, 25 Oct. 2021, https://www.theverge.com/22743753/facebook-tier-list-countries-leaked-documents-content-moderation

SFCG, Search for Common Ground, Handling harmful content online: Cross-national perspectives of users affected by conflict, April 2021, https://www.sfcg.org/wp-content/uploads/2021/07/SearchForCommonGround_Handling-harmful-content-online-report_April-2021.pdf

TBIJ, Facebook accused by survivors of letting activists incite ethnic massacres with hate and misinformation in Ethiopia, Feb. 2022, https://www.thebureauinvestigates.com/stories/2022-02-20/facebook-accused-of-letting-activists-incite-ethnic-massacres-with-hate-and-misinformation-by-survivors-in-ethiopia

UNESCO, Addressing hate speech on social media: Contemporary challenges, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000379

Feminist perspectives on Social Media Councils

Rachel Griffin
SCIENCESPO, PARIS, FRANCE

'Better never means better for everyone…It always means worse, for some.'

— Margaret Atwood, The Handmaid's Tale, 1985

'When they enter, we all enter.'

— Kimberlé Crenshaw, 'Demarginalizing the Intersection of Race and Sex', 1989

Introduction

The overarching question we were asked to consider in these essays is whether Social Media Councils can reconcile private ordering of online spaces with public values. This seems obviously desirable, especially compared to the values that currently dominate online media: surveillance, consumerism and private profit. But 'public values' is a slippery concept, as the Handmaid's Tale quote suggests. Who is 'the public'? What do they value, and what should they? These questions are ideological and contestable. Values that enjoy consensus among elite groups, majority populations or the media don't necessarily serve everyone. Social Media Councils offer one way to constrain corporate power and make platform companies more accountable. But this raises more questions: accountable to whom, and for what? Who will they make social media better for?

As I'm exploring in my PhD research, discrimination and inequality are pervasive in contemporary social media – from content moderation systems which disproportionately censor marginalised groups, to recommendation algorithms which decide whose voices will be heard, and design features that systematically expose marginalised users to hate speech and harassment. If Social Media Councils are to address such issues, we should be asking how they can be designed to redistribute power towards those who are most marginalised within existing institutions, and how their own governance structures could reproduce existing inequalities.

As Kimberlé Crenshaw's work on discrimination suggests, this involves taking politically contentious positions, not just upholding universally-shared values – and it would make life more difficult for some people, like state security agencies or company shareholders. Yet prioritising marginalised communities can also create more inclusive, safe and diverse online spaces that in some sense benefit society as a whole: when they enter, we all enter.

In this essay, I argue that feminist legal theory can guide such an analysis. While it's not possible to do justice here to the diversity of feminist theory, I am influenced by Adam Romero's framing of feminism as methodology: one which is context-sensitive, empirically-informed, and attentive to how power relations are reproduced within legal institutions. In this sense, feminist legal theory aligns with an intersectional approach – not only asking the 'woman question', but highlighting multiple, distinct yet interrelated ways that legal institutions exclude certain perspectives and interests. On this basis, I highlight three analytical moves that I think are characteristic of feminist legal scholarship and suggest how they could be relevant to Social Media Councils.

Elements of a feminist analysis

What are the law's underlying assumptions?

Feminist methods question the perspectives and assumptions built into legal frameworks that are presented as neutral, consensual or objective. How does the law frame problems and potential solutions? What kinds of harms does it recognise? And where is the law silent? As Mary Anne Franks suggests in a feminist analysis of US privacy law, often the harms and concerns recognised in legal and political debate are those that trouble elite and/or majority groups, while harms affecting marginalised communities are not deemed matters of public interest. In the context of Social Media Councils, this suggests several questions that deserve more investigation.

First, many harms and injustices associated with social media are primarily experienced at the systemic or collective level, rather than by identifiable individuals – for example, because ad targeting or moderation software is probabilistically biased against certain groups. Will Social Media Councils be able to recognise, represent and redress such harms? Existing press councils, a major inspiration for Article 19's proposals, generally investigate issues through individual complaints, and focus primarily on protecting individual interests like reputation. This approach would prevent Social Media Councils from investigating or redressing the most consequential decisions platform companies make, such as the design, operation and resources of moderation systems. It also excludes entirely some bigger questions that feminist analyses might highlight, such as whether largely white, male, Western company executives pursuing shareholder profits should be making those decisions in the first place.

Second, what are the assumptions built into the normative framework Social Media Councils apply? This is typically assumed to be international human rights law, which is often presented as an objective or universally-shared set of values to guide platform governance. Feminist theorists, especially those writing from postcolonial perspectives, have questioned the supposed objectivity or neutrality of international human rights norms and institutions. Might this framework bias Social Media Councils towards a liberal, individualistic approach to social media governance, overlooking more structural, material and distributional issues?

It is certainly possible to discuss such broader questions, like the ownership and control of media resources, in human rights terms – but these more collective, structural interpretations of international human rights law are far from universally accepted. If Social Media Councils pursue consensual interpretations of human rights principles that all stakeholders agree on, this will in practice create a 'levelling down' effect where only the least disruptive, most corporate-friendly interpretations of human rights norms win – an outlook which is anything but politically neutral.

Whose voices are excluded?

As well as asking which values are prioritised, feminist legal theorists question whose voices are heard in practice in legal institutions. This is a crucial question for Social Media Councils, since inviting more multistakeholder participation in overseeing platforms is often seen as a way to make social media more inclusive and egalitarian – yet multistakeholder governance structures often fall short of this ideal in practice.

As Brenda Dvoskin shows, while civil society participation has many advantages, it cannot be assumed to offer an objective, unbiased or representative picture of the 'public interest'. Different social groups have different abilities to make their voices heard in multistakeholder fora, depending on factors like

economic and informational resources, popular and media support, and geographic position. This also reflects the preferences of platform companies – who, through funding and partnership programmes, exercise significant influence over civil society and research in this field. Often, these inequalities will track gender, race, class and other entrenched social divisions.

As well as composition and participation, a feminist approach would question what kinds of expertise and authority Social Media Councils value. Current proposals often seem to envisage them as quite technocratic, with authority deriving from legal expertise, in particular in human rights law. Human rights lawyers and academic experts are, to state the obvious, not representative sections of society. If Social Media Councils prioritise legal and academic expertise, they will primarily include people from privileged racial and class backgrounds – replicating the unrepresentative industry they are meant to oversee, which has been linked to biases and blindness towards the particular experiences and risks faced by marginalised users.

Moving forward, we should ask how Social Media Councils can be set up to include a maximally diverse range of voices, and to value lived experiences and grassroots activism as well as legal expertise and prestigious credentials. We should also ask whether stigmatised or marginalised social 'groups, like sex workers and migrants, will have meaningful opportunities to participate. As things stand, they are largely excluded – both discursively, because the interests of unpopular minorities are not seen as part of the general 'public interest', and practically, because they are less likely to have the necessary resources and access. As well as valuing and respecting their perspectives, Social Media Councils should materially support their participation, for example through grant funding. This of course then raises further questions about how to ensure the councils themselves are adequately funded, while remaining independent from platforms.

What are the gaps between theory and practice?

Finally, feminist research often incorporates more empirical or sociolegal methodologies, highlighting gaps between formalist interpretations of the law and how legal institutions actually operate in practice. One aspect of this is the much-discussed tension between formal and substantive equality. I have previously argued that formal protection of human rights in the DSA and other EU social media laws is likely to offer very unequal protection in practice. Similarly, formally equal rights to petition to Social Media Councils are unlikely to benefit everyone equally: people with more economic resources, free time, education and digital literacy will be much more able to press their case. This is another point in favour of a less individualistic approach. Social Media Councils should not only hear individual complaints, but should be empowered to proactively investigate structural and systemic governance issues.

Another related move is to emphasise the socioeconomic context in which legal institutions operate. Here, a particularly relevant question is the relationship between Social Media Councils and platform companies. Feminist sociolegal scholarship has highlighted that where the law delegates responsibility for enforcing the law or policy goals to companies – even seemingly progressive goals, like preventing gender discrimination – they are in practice often enforced in a way that suits corporate interests and demands little change to business practices. This is particularly the case where the law is abstract or focused on procedure, leaving companies to fill in the details.

In this context, given the sprawling complexity of contemporary social media, there is an obvious risk that Social Media Councils will mostly provide rather high-level, abstract guidance on how companies should approach social media governance – which the companies will then be able to interpret in selective and self-interested ways. If Social Media Councils are to operate within the industry as it is

currently configured, dominated by a few large private corporations, we need to ask how they can exercise sufficient power to demand substantive changes in business practices. However, we should also question whether meaningfully inclusive and equal social media governance is possible within this privatised corporate landscape – and whether Social Media Councils could help guide a transition to less commercial alternatives.

Three paths forward for Social Media Councils

On this note, I want to briefly offer some suggestions as to how these approaches could guide us in analysing and potentially constructing Social Media Councils in different contexts. Feminist legal theory can help us analyse the strengths and shortcomings of existing institutions and proposals. It can also help critique and inspire new institutions in platform governance – in both reformist and radical forms.

An existing example: the Meta Oversight Board

Feminist approaches could help us critique – and perhaps improve upon – the most-cited existing example: the Meta Oversight Board. Some aspects of the Board's approach are quite aligned with feminist perspectives, such as its attempts to use selected individual cases and advisory opinions to highlight systemic failings and unequal treatment in content moderation. However, feminist analyses might highlight the Board's unrepresentative composition, which clearly overrepresents Western perspectives and legal expertise. This can be linked both to its rather liberal, individualistic approach to human rights law, tending to prioritise individual freedom of expression over other values, and to its lack of independence from Meta, which continues to oversee the selection of Board members and has vetoed candidates with expertise in targeting, surveillance and other aspects of its business model.

Building on this, feminist analyses might highlight that the Board ultimately functions only within the parameters set by Meta. Its policy recommendations depend on Meta's willingness to implement them. The Board's responsibility is also defined by Meta, as asking whether the company is living up to its own stated values and policies. This means it can't address more fundamental questions, such as whether Meta's business model, surveillance practices, and global influence are conducive to creating an inclusive and egalitarian online environment. Feminist analyses might investigate whether alternative approaches to corporate self-regulation could give more weight to marginalised voices and exercise more leverage over companies.

A reformist project: the Digital Services Act

Second, Social Media Councils could be part of current reformist agendas, like the DSA, which aim to improve accountability within the current ecosystem of corporate social media. Importantly, many DSA provisions – especially very large platforms' obligations to address systemic risks, which Niklas Eder's contribution highlights as a key mechanism enabling regulators to address structural harms and discrimination – remain quite vague and abstract. How they will be enforced in practice is thus still up for debate. As Niklas suggests, civil society and academics can play a crucial role in shaping more progressive and radical interpretations. Through independent research, oversight and recommendations, Social Media Councils could push platforms and regulators to use these provisions to address inequality and discrimination in social media governance, highlighting systemic issues and underrepresented or minority concerns.

However, it is essential to consider how Social Media Councils themselves could become unequal and unrepresentative. How can they be made as diverse and inclusive as possible, which will require

extensive effort and resources, while also remaining independent? Instead of relying on abstract 'public values' which generally reflect majority or elite consensus, how can they enable inclusive political debate and contestation? As the DSA starts being implemented and enforced, feminist legal theory can provide guidance on how to incorporate Social Media Councils into its framework, but will also offer important methodological and analytical tools to investigate how such institutions function in practice and which perspectives and interests they prioritise.

A radical alternative: independent non-commercial governance

Finally, Social Media Councils could be part of more radical structural reform. Arguably, having the online public sphere ruled by profit-making companies – whose priorities are promoting advertisers' brands and selling people things they don't need, not creating safe and inclusive public spaces – is simply not compatible with feminist ideals. Socialist writers like James Muldoon and Ben Tarnoff have argued that governing internet infrastructure in line with public values requires 'demarketisation', with a bigger role for non-commercial and public services. More recently, introducing an explicitly feminist perspective, Rachel Coldicutt has argued for the provision of public resources to replace centralised corporate platforms with diverse online 'counterpublics'. Given the power of today's big tech companies, radical reform might seem distant – but the rapid recent ascent of decentralised non-profit platform Mastodon shows that alternative governance structures are possible.

Could Social Media Councils play a role in governing public and non-commercial platforms? In this context, a key issue is how to ensure that state funding and publicly-owned infrastructure do not translate into direct government control of platforms, which raises obvious issues around democracy and freedom of expression. Delegating governance or oversight functions to councils of civil society representatives could thus provide a path forward for inclusive and independent governance. Such models have successfully been established for older media, such as Germany's Rundfunkräte which oversee public broadcasters. We can learn from such examples, as Bettina Hesse's contribution (link) shows.

Yet this example also reminds us that ensuring that such bodies are inclusive and egalitarian is not easy. The Rundfunkräte remain highly skewed towards white Germans and dominant social segments like Christian organisations, while almost entirely excluding minorities like Muslims and migrants. Considering whose voices would be excluded or included, the international scope of social media also poses tricky problems. If public-service digital infrastructure is provided at the national or European level, is it possible to implement governance mechanisms which do not centre Western perspectives and interests?

Conclusion

If Social Media Councils are to make platform governance better – more rule-bound, more accountable, more democratic – we must ask: better for whom? Feminist legal theory suggests that behind the universalising rhetoric of public values, the default answer is often better for existing elites and privileged groups.

In a worst-case scenario, Social Media Councils might just be groups of mostly white, male, upper- and middle-class lawyers, offering non-binding recommendations on how currently-dominant companies could slightly tweak their procedures and policies to pay lip service to human rights law. On the other hand, if they are given the resources and authority they need, and if their design and operations consistently prioritise the inclusion and elevation of marginalised and minority groups, they could also

push platforms and regulators to address systemic inequalities, make platforms safer and more inclusive spaces, and establish more egalitarian forms of social media governance.

Achieving this will not be easy and will require critical analysis and concerted efforts to excavate both how women and minorities are excluded within contemporary social media, and how they might continue to be excluded within new governance institutions. In this essay, I have tried to identify some first questions that are suggested by feminist approaches to legal research, and which deserve further investigation as proposals for Social Media Councils move forward.

Learning from Broadcasting Councils

Bettina Hesse
VER.DI – UNITED SERVICES UNION, BERLIN, GERMANY

When considering the design of platform councils, it makes sense to look at possible role models. Independent institutions whose task it is to control and align with society communication spaces free of state interventions have existed for a long time: the twelve public broadcasters in Germany are supervised by broadcasting councils. For 70 years these boards have been meeting to discuss programmes, elect management and decide whether to release funding for new projects. Even though according to the Federal Constitutional Court, they are a very crucial component of the independent, non-governmental public service system, they operate mostly independent of the public and have barely been known in society. This changed abruptly when an alleged corruption scandal was uncovered in the summer of 2022, in which the chairman of the board and the director of the Berlin-Brandenburg Broadcasting Corporation were suspected of being involved. The following article explores the question of which conclusions can be drawn from the current crisis in German broadcasting supervision in order to set up platform councils more efficiently than the broadcasting councils have proven to be.

Social supervision as condition for independence

One of the constituent features of the publicly funded broadcasting system founded after WW II in Germany has been social control – with the aim to prevent political instrumentalisation and to foster democracy instead. Broadcasting is supposed to be free of any state interference and has to provide relevant contributions to everyone in society. In order to align it with public values, the responsibility for the journalistic-editorial mission of the public broadcasters has to be carried by the society. In 1961 the Federal Constitutional Court stipulated that society should be represented in the broadcasting councils by members of all significant political, ideological and social groups (appointed by the respective Federal State). According to the Federal Constitutional Court, they would serve as the "trustees of the general public interest" (1961) and constituted the "highest organ of the institution" (1971). This ruling was prompted by the observation that political representatives were increasingly dominating the oversight bodies. Despite the verdict, this tendency persisted and led to another court decision in 2014 that asked the councils to limit the amount of governmental and administration representatives to one third at maximum.

The composition of each broadcasting council is determined in different ways, depending on the legal basis in the respective Federal State. The State parliaments can send a certain number of members to the councils. In addition, organisations and associations are appointed, which may also send representatives. These are usually religious groups, employers' associations, trade unions, and in some States social associations, LGBTIQ*-groups, environmentalist groups etc. In addition, there are sometimes "citizens' mandates" for which citizens can apply to the State parliament, and in some cases, the broadcasting council can elect additional representatives.

Broadcasting Councils

The main task of the broadcasting councils is to review the fulfilment of the public service mandate in the form of ex-post control (decision-reviewing) – without intervening in the program, because the freedom to shape programming lies with the broadcaster's director. The council has to monitor legal

requirements in the programming, such as standards for the protection of children, respect for human dignity, freedom of expression and religion, freedom from discrimination, gender equality, the representation of existing political opinions etc.

Furthermore, the broadcasting council elects an administrative council, which has the task to supervise the finances of the broadcaster and to permit larger expenses (starting from 30.000 to 5.000.000 Euros).

In addition, the broadcasting council elects directors, suggested by the administrative council. Except for the election of the director-general and other directors, the broadcasting council has no right of initiative, yet almost all far-reaching measures taken by the director-general depend on the approval of the broadcasting council or the administrative board.

Since 2009 the councils also decides upon the introduction of new digital services by assigning "three-step-tests" – an assessment of the new service's contribution to the fulfilment of the given constitutional task.

The members of the broadcasting and administrative councils are volunteers. However, they receive expense allowances, meeting allowances and travel allowances – in very different amounts. In addition, depending on the broadcaster, a council member receives an amount starting from a fixed rate of 264 Euros per year (Radio Bremen) to 1.000 and up to 2.800 Euros per month (WDR). The councils receive (more or less) support from an office (with 0.5 to 7 FTE), which is sometimes equipped by the broadcasters, other councils can decide on the financial and personnel resources of their offices themselves.

Criticism and design issues of Broadcasting Councils

Systematic comparative studies on the staffing, working methods, decisions or the intra-organisational division of labour between broadcasting council and administrative council are rare in German research. Also in the public debate on public broadcasting and in the legislative process of structural development, the supervisory bodies play at best a secondary role. Therefore, the following remarks refer to a debate of a usually small professional public.

Various points of criticism can be extracted from the academic as well as the media debate. They indicate doubts on whether the councils actually fulfil their function or whether there are issues with their design:

Influence of political parties

Despite the 2014 verdict, 41% of the members in the council of "Deutsche Welle" belong to government or administration. With at least 18%, state representatives form the largest group in every council. Members with political affiliation tend to have strong influence in the councils due to the resources provided by their sending organisations, their respective expertise and opinion leadership and their degree of organisation within the councils ("Freundeskreise"). As a result, there is an imbalance of power and representation between the political and civil society councillors.

Representation in the councils

According to a study by "Neue deutsche Medienmacher*innen" the power relations prevailing in society are perpetuated in broadcasting councils: The socially marginalized (poor or disabled people, religious minorities, PoC, LGBTIQ*) are underrepresented, the average age of the members is relatively high (57.8 years) and more men than women or persons of other gender are members in the councils. This

gives rise to concrete problems, since the broadcasting councils monitor compliance with the programme mandate. Those programme standards typically include respect for human dignity, freedom from discrimination or gender equality – requirements whose non-observance particularly affects disadvantaged social groups. In case of contentious issues in broadcasting, it would be helpful to hear experts and stakeholders of these groups in the board, especially if they are not represented in the board themselves. The above-mentioned study points out that in most broadcasting councils such hearings either do not take place at all or only very rarely.

Associations without a seat on the broadcasting council often complain that the allocation process for seats on the council is not transparent and that there is no broad public consultation in the legislation process. With the last amendment to the WDR Act, achieving diversity in the broadcasting council was made even more difficult by the deletion of citizens' mandates. In contrast, the number of seats for state representatives remained untouched.

It is often argued that the broadcasting councils represent outdated social relations. The majority of councillors are chosen by political parties and associations – whose role in society has changed considerably over the last 75 years. Today, professional communities of users hardly organise themselves as classic interest groups. Therefore, young and digitally skilled people are structurally underrepresented in all councils.

Alignment between broadcasters and the public

Criticism is often raised that broadcasting has distanced itself from society more than a publicly funded system should (although representative surveys show a high level of general trust in public broadcasting). This could be interpreted as criticism in the broadcasting councils, since discussing the broad lines of the programme with the broadcaster is their key task (with the broadcaster holding the program design authority). In fact, many people contact the broadcaster with submissions, complaints and suggestions about the programme. However, submissions are answered by the directorate. Only if complainants are not satisfied with a reply, they can demand that their complaint is dealt with by the broadcasting council. Only then does the council hear about the audience's criticism. There is no institutionalized dialogue between the council and the public.

High engagement indicates a sense of ownership and voice, which certainly originates in the fact that all citizens in Germany (have to) finance broadcasting. Regarding the councils, the results of the survey #meinfernsehen2021 express the audience's strong desire for reform and change – even though it was not always clear to the participants how the supervisory bodies work, what exactly the thematic focus of their work is and what their relationship to the broadcasters is.

Transparency of the council's work

Since the beginning of the Covid-19 pandemic there have been livestreams for every council's full sessions, yet not from the meetings of the committees (or of the administrative boards), so the process of decision-making is not accessible to the public. The agendas are public, but difficult to find, and the minutes of meetings are not sufficiently informative. When important decisions are made or personnel changes, the councils (often: the broadcasters) publish press releases (sometimes only one release per year). Usually there is no media coverage of the meetings and only a few councils use communication instruments such as newsletters. All in all, the councils mostly act with distance to the public.

Independence from the media organisation

For many decisions the councils depend on information provided by the house's director, the house's legal department or the council's office (usually equipped by the house). No clear rules are communicated to most councillors on whether they are even allowed to commission external expert reports themselves, which is therefore a rarely used instrument.

Expense allowances are paid by the broadcaster, which certainly does not contribute to a sense of independence from this institution.

Self-concept of councillors

Some members of broadcasting councils identify as inspectors, whereas others are accused to act as co-managers of the broadcaster. The size of some councils (from 26 to 74 members) may contribute to diffusion of responsibility and low commitment of individuals.

Qualification and volunteering

There has been a long debate whether the broadcasting or administrative councils should consist of (more) media/law/accounting experts instead of representatives of the broad public. It can hardly be questioned that expertise is required for effective control. However, defenders of the pluralism model argue that expertise can be acquired on the job – as long as the term of office is long enough and the members find the necessary time for further training in addition to their main job. The concept of volunteering is in question as well, since - including all committee sessions - there can be up to 34 meetings per year – which is difficult to combine with a full-time job (and explains why many of the members join the councils when retired).

Continuity and identification

Many councillors report that it takes several years to be able to participate effectively in the council. When new members come into office every 5-6 years – or even earlier, in the case of shared mandates – the council's monitoring work lacks continuity. On the other hand, there are demands for term limits to prevent too much identification with the respective broadcaster.

Lack of flexibility

Media use and production have changed considerably since the councils were first introduced seven decades ago. The shortened production cycles of cross-media content require forms of handling and control which the old structures cannot react to adequately. Some, but not all councils have established telemedia committees, but no bigger changes have been introduced.

Efficiency

The efficiency of the control of broadcasting councils has been contested for a long time, as some councils had the reputation to just pass uncritically what the director suggested to them. A recent supposed case of nepotism including the director of one of the public broadcasting stations (RBB) and the chairman of its administrative council exposed the system's shortcomings quite obviously.

A clear lack of accountability was revealed: the members of the administrative council each worked alone, their work was not sufficiently transparent to the other members. This resulted in a lack of mutual

checks and balances, which apparently was exploited at least by the chairman and led to serious mismanagement.

Both broadcasting and administrative council realised that the decision papers they had received from the broadcaster's directorship and on which their decisions depended, were not as neutral and complete as assumed. The council's office apparently was not able to generate independent information. There was no legal base for the council or its office to commission external appraisals. And it turned out that the broadcaster's employers had received explicit instructions to withhold information from the administrative board.

Investigations around the incidents are still in progress, and it is well possible that personal misconduct played a bigger role. Nevertheless, these findings indicate shortcomings in the design of broadcasting councils. The fact that the irregularities at RBB have been revealed by journalists from a commercial news outlet and not by members of the internal supervision bodies speaks for itself.

Current legislation: more responsibility for Councils

In reaction to this scandal, the federal state of Hessen just passed a law, which is supposed to guarantee the independence of the council's office from the broadcaster.

In the next months, a new amendment to the Interstate Treaty on Media will be debated in the state parliaments. The amendment addresses some of the previously mentioned, long known concerns regarding the broadcasting councils. The new amendment to the Interstate Treaty on Media will grant them more responsibilities. Councils will no longer be limited to discussing the programme in the aftermath, but they will also set up quality standards for the programme in advance – a shift to rule-making. Furthermore, their responsibility in financial control will grow. They will review compliance with the principles of economy, efficiency and financial rigour, and monitor resources efficiency, which was previously the task of the internal revision as well as that of the commission for determining financial requirements in broadcasting (KEF). Given the recently disclosed control deficit in the RBB and the already known weaknesses in the board-design, these new tasks will be an enormous challenge.

The Interstate Treaty also provides for more dialogue between broadcasting stations and audience – a notable decision, given that the already existing broadcasting councils have been designed as a forum to align the programme with the audience. The legislator reacts to an increased need for participation in public broadcasting. A large number of dialogue formats between broadcasters and audience have been instituted following this widespread demand in the past years (online-panels, townhall-meetings, public debates, open-door days, focus groups for the further development of broadcast formats, WDR "Userlab" for testing digital applications) – without a legal mandate. In addition, accompanying research projects (like #meinfernsehen2021, similar to ARD-Zukunftsdialog 2021) gathered feedback on the quality of and ideas for public service broadcasting.

Leveraging experiences for supervision of new platforms

If the broadcasting councils are to contribute to a sense of ownership of the public, then public awareness of this institution and its work is a necessary prerequisite. It is therefore obvious that the main problem of the broadcasting councils is that the public knows far too little about them, their work and its impact. The very existence of broadcasting councils seems not to bring the intended legitimacy (anymore), there seems to be a higher demand for explanation and participation and an actual need for stronger societal control. If broadcasting councils are to serve as inspiration for Social Media Councils, the following

conclusions from experiences of broadcasting councils should be considered to improve the design of new supervisory bodies:

Legitimacy and alignment through visibility and transparency

- Transparency of the council's work, of consideration processes and of the impact of the council's decisions is necessary. The council's attitude towards the public must be one of openness. It has to communicate actively in order to raise public awareness for its existence and its work.
- A dialogue between council and audience should be established – as an occasion to explain the structures of the council and its work and to learn about the audience's concerns. The audience/users must be able to contact the council directly.

Adequate representation and accessibility

- A continuous process to guarantee the representation of an evolving society should be enabled. The general public should have the chance to participate in the composition process of a council.
- In order to prevent political dominance, it should be considered to include more social groups on a parity basis instead of proportional political representation that leaves room for fewer groups.
- Sending organisations should be encouraged to consider members of different ages, backgrounds, religions, sexual orientation and gender identities and disabilities when selecting their representatives. Previous experiences concering gender equality show that only binding regulations work.
- Representation deficits of people without affiliation could be reduced by partly random selection of council members, by publicly elected members or by institutionalised dialogue formats that are promoted and open to the public. Other possibilities would be to oblige political parties (or equivalents with fixed seats) to select representatives from a pool of defined organisations or to publicly advertise their positions.

Quality and independence

- When sending delegates, organisations should prioritise relevant expertise over prominence of individuals.
- Councils should be provided with support in form of independent and reliable information/preparation of decisions. The council must have access to independent media research. Ensure that the council has the legal competence to commission expert opinions itself.
- Members should be provided with resources to fulfil their task with the necessary diligence. Civil society members may need paid leave from their main job in order to adequately fulfil the tasks of the council and for trainings. Financial compensation must be based on their needs.
- Allowances should be paid by an independent entity, not by the supervised organisation.
- Committees/discussion forums should consist of a small number of people so that everyone has their say and feels commitment.

Platforms considerably differ from public service broadcasting in many aspects. However, since there are decades of experience – and criticism – with broadcasting councils, it might be worth taking note of these supervisory bodies when designing social oversight structures for new communication spaces. Undoubtedly, learning from broadcasting councils, the design must be carefully crafted to provide for

efficiency, transparency, independence and representation, so that councils can really improve alignment with the public interest and contribute to an infrastructure's legitimacy.

Assessing the systemic risks of curation

Niklas Eder
OVERSIGHT BOARD, LONDON, GREAT BRITAIN[172]

The future of content moderation (self-)regulation is systemic

Systemic risk assessments are the next big thing in content moderation. They carry a gigantic promise, which is to address content moderation where it really matters. If things go well, they will force platforms to engage in the "proactive and continuous form of governance" that we have been waiting for. Systemic risk assessments, so the hope, will constitute the second core pillar of content moderation (self-)regulation, and overcome the limitations of individual remedy mechanisms, which constitute the first pillar. The amplification of content, the design of recommender systems, the newsfeed — here referred to as curation practices — are out of reach for individual remedy mechanisms, yet are believed to be responsible for the deterioration of civic discourse, electoral processes, public health, security and fundamental rights. Articles 34 and 35 of the Digital Services Act (DSA) come to rescue, and finally establish obligations for platforms to assess and mitigate these risks, ideally with participation of stakeholders. These risk assessments are then sent to auditors, the Commission and the Board for Digital Services, which will review the risk assessments, develop best practices and guidelines, and will potentially require platforms to take alternative mitigations measures or even fine them. This is huge leap forwards for content moderation. It's also a jump in the unknown.

The compelling idea behind systemic risk assessments is that, to mitigate the harms of platforms, we need to go beyond individual remedy. The great challenge, which this approach entails, is that we have no idea how to measure and mitigate systemic risks of content moderation. The DSA vaguely defines the sources of risk (Article 34 § 2), the kinds of risks (Article 34 § 1) and the kinds of mitigation measures platforms must consider (Article 35). Beyond that, systemic risk assessment in content moderation are unknown territory for industry, academics, and regulators alike. We need to make systemic risk assessments work to address the societal harms of content moderation, but we don't know yet how. The next months offer a unique opportunity for academia to help transform systemic risk assessments into the effective mechanism they could be, rather than the harmless paper tiger they might become.

A virtuous loop to assess systemic risks

Beyond defining sources of risk (Article 34 § 2), kinds of risks (Article 34 § 1) and the mitigation measures platforms must consider (Article 35), the DSA outlined responsibilities and processes around systemic risk assessments. It basically creates a loop in which, once a year, platforms efforts to assess and mitigate risks are evaluated and recalibrated. Platforms are obliged to work with stakeholders and assess risks and develop mitigation measures, and to submit a yearly report on their efforts to the EU Commission, the Board for Digital Services and auditors. They will then assess whether platforms comply with their obligations under the DSA, or whether they should alter their risk assessment efforts and take alternative mitigation measures. After the Commission and the Board for Digital Services

---

received the first round of risk assessments from platforms, they will begin developing best practices and guidelines which platforms, in their future risk assessments, will be expected to apply. While the setup of this process is reasonable, it raises as many questions as it answers. The success of systemic risk assessments ultimately depends on who takes what role in that loop, and how those involved execute their particular functions.

Neither platforms nor EU institutions are particularly well equipped to decide over what systemic risk are and how they should be mitigated. Platforms are ill equipped, because self-assessments by corporations whose ultimate goal is to maximise profits, lack credibility. The Commission and the Board for Digital Services are ill equipped, because public institutions should not dictate the rules governing content moderation and with that public discourse. It also appears rather unlikely that auditors will play a significant role in shaping systemic risk assessment. Conventionally, and this matches the description in the DSA, the role of auditors is to assess whether a company has complied with legal requirements. Auditors are usually large companies, likely with no significant expertise in content moderation and certainly with no particular legitimacy with regard to defending the public interest. The cannot set the standards governing systemic risks.

This leaves one more role open, on which we may rest our hopes, which is the role of stakeholders. This role is only established in the recitals of the DSA, indicating minor relevance. However, taking into account the practical and theoretical concerns raised here, stakeholders might be best equipped to shape the standards based on which systemic risks should be measured and mitigated. The Commission could easily empower stakeholders, if it decided to give their contributions significant weight in its evaluations of the submissions of platforms, and if it reflects them in their best practices and guidelines. Perhaps this is the true ingenuity of the DSA's systemic risk assessment provisions: They create a loop in which the Commission can use its enforcement powers to empower stakeholders. This virtuous loop might solve the conundrum of public actors regulating speech: It allows them to assure that platforms will account for public interest, without public actors having to define what that concretely means themselves.

The hard question, which will be addressed in the remainder of this essay, is what shape stakeholder engagement should take, if it is them who are supposed to develop the standards based on which to assess and mitigate risks. Simple stakeholder consultations, conducted by platforms, might have little effect. Much speaks for a more substantial cooperation between platforms and stakeholders. We can think of a variety of constellations, many of which already exist to a certain degree. For example, platforms can — breathe in, breathe out — work with Social Media Councils to establish the standards which govern systemic risk assessments.

Social Media Councils

The cooperation between platforms and Social Media Councils can take different shapes and different degrees of intensity. Fuelled by the DSA, different models could emerge.

On one end of the spectrum could be cooperations which leave control largely in the hands platform, allowing them to decide what questions the Social Media Council should answer, what access they get to information and leaving them discretion if and how to implement solutions suggested by the council. On the other end of the spectrum, the cooperation could consist in platforms conveying certain competencies to Social Media Councils, providing them with access to relevant data, allowing them to chose the questions they wish to engage with, to develop their own frameworks and empowering them to make recommendations on some of the most important questions pertaining to curation and systemic

risks. The platforms would not necessarily have to follow all recommendations that Social Media Council make, but at least would commit to respond to them — and justify before the Commission and the Board for Digital Services if they did not follow the recommendations the Social Media Councils made. The Social Media Councils would in turn work with relevant stakeholders to reach their recommendations, ultimately strengthening the impact all stakeholders have on the focus and outcome of systemic risk assessments.

Relying on Social Media Councils has significant advantages for all parties involved, including platforms and the Commission: Platforms would benefit from involving Social Media Councils in their process of assessing and mitigating risks, as Social Media Councils would ultimately help decide the hardest questions platforms face. If platforms chose to implement the recommendations by the Social Media Council, and incorporated them into their strategy to assess and mitigate risks, this conveyed credibility to the systemic risk assessments they submit to the EU. In response, platforms could expect positive evaluations, non-interference from the side of the EU institutions and would minimise the risks for fines. The involvement of Social Media Councils allows the Commission to restrain from defining the concrete standards against which to measure content moderation and curation practices, as any public institution should, while catalysing a productive, a virtuous loop which empowers civil society. Thorough the platforms' work with external experts and Social Media Councils, the Commission and the Board for Digital Services receive high quality risks assessments, based on which they can develop practical and effective best practices and guidelines without itself unduly interfering with the working of platforms.

Empowering positive data rights with platform councils

Dominik Piétron
HUMBOLDT UNIVERSITY BERLIN, BERLIN, GERMANY

Every social interaction on the Internet, whether in social media, e-commerce, or digital public services, presupposes a certain form of data collection, data storage, data processing and data transfer. However, data governance practices mostly remain opaque. Especially the most influential data intermediaries, digital gatekeeper platforms such as Google, Facebook, or Amazon, are autocratically ruled by a non-transparent managerial elite. Numerous court cases have highlighted how the platform's data governance can be significantly harmful to individuals and groups, workers, and businesses, as well as the public discourse. Nevertheless, users are increasingly disclosing very intimate personal information about their behavior, social relationships, needs, and desires, thus boosting the data power of platform companies. This central contradiction between informational self-determination and algorithmic heteronomy illustrates the blatant lack of fundamental democratic principles in the platform economy.

In this article, I argue that rights-preserving data governance that adequately addresses the individual and social harms caused by digital platforms requires a new institutional form for collective decision-making. Given that every platform operator is confronted with the need to govern personal data, democratic platform councils are a promising institutional framework for many areas of the digital society – not only for social media platforms, but also for e-commerce and public service platforms. Just as English entrepreneurs in the 16th century demanded a representative parliament to decide how to spend their tax revenues, and workers in the 19th century fought for the right to establish works councils, today's platform users require some sort of platform council to debate on the data governance mechanisms and enforce their data rights.

I will proceed in four steps: First, I will elaborate on the need for democratic data governance on a collective level that originates from the individual right to informational self-determination. Second, I argue that platform councils are a well-suited institutional framework for democratic data-governance. Third, I develop design principles for platform councils that form the minimum conditions for empowering data rights of individuals and collectives. Fourth, I empirically illustrate my argument using the example of "Berlin.de," a German city platform that acts as a central digital access point to Berlin's public infrastructure.

The need for democratic data governance

The challenge of data governance arises, inter alia, from the paradox that the right to informational self-determination is legally enshrined in the EU with the GDPR but can hardly be implemented in today's platform economy. There are two reasons for this, which can be explained along the vertical-horizontal-axes of data relations developed by Simone Viljoen (2021: 607): First, there is such a strong vertical asymmetry of power between users and platform operators that user's consent to processing of personal data is often coerced rather than voluntary (ibid.). Due to the strong concentration tendencies in the platform economy, some platforms like Alphabet's and Apple's smartphone OS and app stores, Amazon's e-commerce empire, Microsoft 365 or Meta's social media group have become essential infrastructures of our everyday life and important preconditions for social participation. But sector-

specific platforms such as AirBnB or Uber are also increasingly displacing analog services and are already without alternative in some regions of the world. The problem is that users must agree to far-reaching data usage declarations when they sign up for platforms, thus allowing platforms to gain "quasi-ownership" through the enclosure of personal data. As the digital socialization of the individual extends to more and more areas of life, the individual is severely overburdened with a rapidly increasing number of decisions to be made about the processing of personal data.

In addition to vertical power asymmetry, a second reason for the inadequacy of individual contract-based data rights, according to Viljoen, is horizontal data relationships between different users (ibid: 609). Horizontal data relationships occur automatically when many people are given a digital identity that tracks their behavior, records it in a digital form, and aggregates it in a database with other data subjects. As people are assigned different data values, relative differences become visible and social patterns respectively horizontal data relations emerge. Individuals can be correlated and categorized into groups, enabling population-level insights and predictions that form the basis of most data-driven value creation – as well as many data-driven social harms such as algorithmic discrimination, disinformation, or behavioral control. The crucial point is that the sharing of data by one person in a group also allows conclusions to be drawn about all the other people in the group. As social entities, humans are always affected by the data decisions of their friends, peers, neighbors, coworkers, etc. This is most obvious in communication between users, where the message becomes public as soon as only one user shares the data. Thus, individual data reflect social relationships with others and thus represent an inherently collective resource that cannot be reduced to individual choices.

Consequently, the individualistic approach of the GDPR is increasingly criticized as insufficient. There are calls from various sides to treat aggregated personal data as "data commons" (Bria 2018, Shkabatur 2019, Zygmuntowski et al. 2021), that belong to a community and therefore require a collective governance framework – just as Nobel laureate Elinor Ostrom has emphasized the commons nature of natural resources such as forests, waters, or traditional knowledge (Ostrom 1998). Singh and Gurumurthy (2021) even propose an additional legal codification of collective data rights. They hold that the rights to determine the use of a social group's data, be it the users of a digital platform or the residents of a city, should belong to the respective community as a collective subject. But even without the demand for collective data rights, a superindividual data governance is necessary, since the "bundle of data rights" (Kerber 2021) must be assigned among various actors, that include rights on non-personal data as in the case of IoT products. Either way, a supra-individual decision-making process is required, in which the rights of the individual, such as the right to data portability or the right to be forgotten, fully exist and are supported. Starting from such a rights-based data governance approach requires a collective decision-making process in which all individuals decide collectively on the use of their aggregated personal data. Since that can only be legitimized by the aggregated will of those affected by data processing, Viljoen concludes that a "democratic data governance" (Viljoen 2021: 648) is needed to enforce informational self-determination.

Platform councils as an institutional form for democratic data governance

Since platforms are technically algorithmic infrastructures for controlling data flows, data governance is at the heart of every "platform governance" (Gorwa 2019). Each digital platform has established its own data governance structure, that can be defined as the sum of decision rules, rights and obligations for the collection, storage, processing, and sharing of data within an organization and between an organization and external actors (Abraham et al. 2019, Khatri/Brown 2010). As information brokers it

is their very function to reduce transaction costs by deciding on the adequate social embedding of data flows. By specifying a data governance framework, platform operators have to give answers to a multitude of difficult data questions: Which product should be listed first in the search results? At what point is a social media post no longer acceptable? How should algorithms standardize social interaction in the digital space? How much personal information is necessary for a service to deliver the best user experience?

Numerous scholars have highlighted the legitimacy deficit of platform operators, especially regarding oligopolistic decision-making at big digital platforms such as Amazon, Google, Facebook, or Twitter (Griffin 2022). While on the one hand European data laws such as the Digital Services Act or the Digital Markets Act ensure that platforms follow rule of law principles, on the other hand the idea of "multistakeholderism" is emerging to increase civil society's influence on platform governance through transparency, consultation, and participation (ibid). In this context, platform councils have recently become a normative approach through which platforms renegotiate their relationship with their users and demonstrate their commitment to public interests. Especially in the field of social media platforms, the first platform advisory boards were established, such as Meta's Oversight Board or the advisory boards of Twitter, Tiktok, Twitch and Spotify (Kaye 2019, Kettemann/Fertmann 2021).

There are good reasons why platform councils could be a useful addition to social media platforms as well as e-commerce and public service platforms. Numerous court cases have shown how data management on e-commerce platforms like Amazon can harm individuals and groups, workers and businesses through its power to select and amplify certain data objects. The same applies to public service platforms, which often hold a local monopoly and thus, in the course of digitization, act as a central information provider for a city and increasingly moderate content.

Beyond that, however, platform councils could be an appropriate institutional form for democratic data governance, which necessarily arises from the tension between data rights and algorithmic heteronomy. In the platform economy, it is the rule that users de facto give up their rights when they agree to the terms and conditions of a gatekeeper platform and have no further say in the processing of their data. This is particularly problematic because many users are tied to specific platforms for which there are no alternatives, and platform operators can easily reprogram the platforms' algorithmic infrastructure, leaving users with the choice of accepting the new terms or losing access to the platform. Bringing platform governance in line with European data rights would thus mean putting informational self-determination at the center and understanding them not just as a negative right to protection, but as a positive right to co-design data infrastructures. To this end, a collective decision-making process is needed that gives platform users a say in the design of the platform architecture that channels their data. Platform councils could strengthen this positive data right and overcome the two main shortcomings of the General Data Protection Regulation: They are collective rather than individualistic, so that horizontal data relationships are considered, and they are dynamic and process-oriented rather than static contractual relationships, so that data rights are also taken into account as platforms evolve.

Design principles for positive data rights-based platform councils

The crucial question then is, how does a platform council that supports a democratic data governance look like in practice? When it comes to designing governance mechanisms for collective goods, Elinor Ostrom's work on the institutional diversity of natural commons all over the world is a valuable point of reference. Especially Ostrom's approach to develop certain "design principles" (Ostrom 2010: 652) as opposed to a one-size-fits-all model can be of great help here. Ostrom starts with a resource to be governed collectively and a community of people who have a stake in the resource, since they are

producing or using it or are somehow affected by its usage. Applied to the challenge of data governance in the platform economy, the resource in question is aggregated personal data, and the community boundaries can be constituted by platform users as well as external individuals affected by decision making based on the aggregated personal data of the platform. Based on this assumption we can derive the following design principles for data rights-based platform councils in various forms:

Inclusive discussions and feedback opportunities

All platform users – but also non-platform users affected by data governance of a platform – should be able to participate in an inclusive discussion about how user-generated data is collected, aggregated, analyzed, shared, and fed back to users. In addition to the primary use of data on the platform, it should also be decided in particular which data can be made available for secondary use by third parties. The agenda-setting should be open and autonomous so that every governance mechanism can be debated. This requires an institutionalized discussion forum in which proposals for change and concerns can be raised and discussed. To organize this process, the platform council should select moderators to facilitate the discussion and develop individual proposals based on stakeholder comments.

Democratic accountability

To ensure democratic accountability, clear guidelines need to be established about who decides on proposals and how – from a consensus democracy approach, in which all stakeholders must agree, on the one hand, to a more representative approach, in which a board of elected representatives decides on new governance mechanisms with a relative majority, on the other hand. Given the complexity of data governance issues, platform councils should act as a coordination body that organizes this decision-making process. The more decisions the platform council makes, the more important it becomes to legitimize its composition and to hold it accountable through democratic elections. In free, fair and secret online elections, the various stakeholder groups of the platform (including platform users, platform workers, and external stakeholders) should select their representatives for the platform council. If appropriate, quota can ensure representation of individual groups or minorities. Elections should be held at regular intervals every one to five years so that citizens can hold council members accountable for their decisions.

Algorithmic transparency

Since platform infrastructures often act as algorithmic black boxes, transparency of algorithms and data governance practices is key for a rights-based data governance. Only if stakeholders know, how their digital identities are constructed and how their online and offline behavior is affected by the algorithms of the platform, they can perform their right on informational self-determination. To verify the implementation of the resolutions, the platform council must have comprehensive rights of access to information vis-à-vis the platform operator's technical infrastructure. Platform councils shall contribute to informed decision-making of stakeholders by highlighting potential risks and clarifying trade-offs among competing interests of stakeholders.

Effective implementation

It must be ensured that the decisions of the Platform Council are effectively implemented. To this end, the councils must be technically and legally independent of the platform operators in order to enforce democratically made decisions against the will of the platform operator if necessary, as research on data

trusts has shown (Delacroix and Lawrence 2019). The independence of the platform council must also be ensured by means of a secure financial basis, whereby funding can be provided either by platform operators or by publicly funded media regulators. Furthermore, the voluntary commitment of the platform operators to implement the decisions of the platform council may not be sufficient. If a platform operator does not cooperate, platform councils must have the ability to seek assistance from data and consumer protection authorities to hold the platform operator accountable.
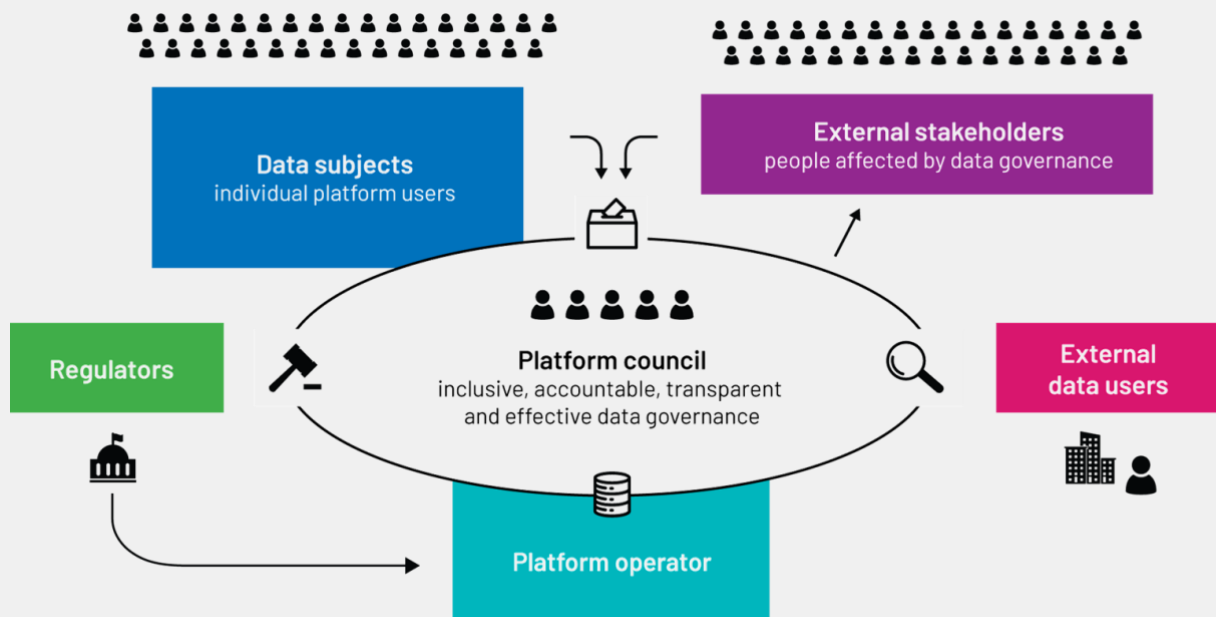
Graph 1: Visualization of a platform council based on the design principles described above.

The case of city platform 'Berlin.de'

While platform councils are currently being established primarily on social media platforms, similar institutions are just as relevant for e-commerce platforms or public service platforms. Public service platforms are also playing an increasingly important role in the digital accessibility of public infrastructure. The behavioral data generated in the process is an important feedback mechanism for administrative modernization and should at the same time be publicly accessible wherever possible, but it also harbors risks and must be protected in certain areas. Also, the question of what personal data should be collected and how it should be processed is not trivial and should be developed in the spirit of informational self-determination, at best with the participation of the users themselves. In the following, the proposed data governance structure based on platform councils is illustrated with the case of "Berlin.de", the central public city platform of Berlin. Recently, the public administration has initiated a process to introduce a platform council, and now develops its own framework to institutionalize collective decision-making in the context of an infrastructural city platform.

The platform Berlin.de was launched in 1995 and successively expanded to become the central digital access point to Berlin's e-government services, providing information on urban infrastructure such as mobility, culture and events, tourism and travel, shopping, and business. It also includes the Berlin Open Data portal and the participation portal mein.berlin.de, which makes the platform a complex digital ecosystem used by various user groups such as citizens, public servants, local businesses, and tourists.

In 2021, the platform was bought by the municipality and turned into public property. In this way, a process was started that continues to this day and, inter alia, led to the idea of a platform council. Already during the communalization, the Berlin House of Representatives decided to do a thorough relaunch of Berlin.de, including a central focus on data protection and inclusive platform development together with local civil society organizations (Abgeordnetenhaus Berlin 2021). Shortly before, several civil society organizations organized in the Bündnis Digitale Stadt Berlin had already published an open letter in which they proposed a participatory process for co-designing the city platform (Bündnis Digitale Stadt Berlin 2021). In the following month, the civil society alliance organized an expert workshop and a public discussion event together with public servants, where the demand for an own platform council was formulated for the first time. In summer 2022, the city government agreed to start a process to develop a platform council for the city platform Berlin.de. A first internal meeting with representatives from the administration and civil society took place in September where both sides agreed to the need of platform councils as a tool for participatory platform-co-design and collective data governance. In October 2022, the common goal of a platform council for Berlin.de was re-affirmed in the House of Representatives (Abgeordnetenhaus Berlin 2022).

However, the concrete institutional form of the Platform Council Berlin.de has not yet been determined. The House of Representatives in October 2022, and the Alliance for the Digital City of Berlin proposed three guidelines for the design of the city's administration, which are largely based on the design principles developed above in section 3:

– The Platform Council should act as an interface between the administration and the population. It should be a tripartite multi-stakeholder body composed of elected representatives of the various parties of civil society, of the citizens of Berlin and of the public administrations involved, thus combining elements from direct and representative democracy.
– The main objective of the Platform Council is to formulate recommendations for the administration on the development of planned or ongoing projects related to the city portal. To this end, its agenda-setting may not be restricted to certain topics but be open able to respond to users' needs.
– For the users to participate and hold their representatives accountable, the advisory board should meet in public, obtain information from the administration, and conduct its own surveys of platform users.

Sources

Abgeordnetenhaus Berlin (2021): Das Stadtportal berlin.de in öffentlicher Hand neu aufstellen, Nr. 2021/81/42 Drucksache 18/3834, https://pardok.parlament-berlin.de/starweb/adis/citat/VT/18/PlenarPr/p18-081bs3240.pdf.

Abgeordnetenhaus Berlin (2022): 13. Sitzung des Ausschuss für Digitalisierung und Datenschutz am 05.10.2022, Youtube, https://youtu.be/USRCs54JxJY?t=2381.

Bündnis Digitale Stadt Berlin (2021): Info-Sheet „Rekommunalisierung des Stadtportals Berlin.de" – Geschichte, Technik, Herausforderungen, https://digitalesberlin.info/info-sheet-stadtportal/.

Bria, F. (2018): New Deal on Data, London/New York: verso.

Sylvie Delacroix, Neil D Lawrence (2019): Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance, *International Data Privacy Law*, 9 (4): 236–252, https://doi.org/10.1093/idpl/ipz014

Gorwa, Robert (2019) What is platform governance?, Information, *Communication & Society*, 22(6): 854–871, https://doi.org/10.1080/1369118X.2019.1573914.

Griffin, Rachel (2022): Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability, http://dx.doi.org/10.2139/ssrn.4190500

Kaye, David (2019): Social Media Councils. From Concept to Reality, Conference Report, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf

Kettemann, M./ Fertmann, M. (2021): Platform-Proofing Democracy. Social Media Councils as Tools to Increase the Public Accountability of Online Platforms, Friedrich Naumann Stiftung (eds.), https://shop.freiheit.org/#!/Publikation/1084.

Kerber, W. (2022): Specifying and Assigning "Bundles of Rights" on Data, in: Hofmann/Raue/Zech (ed.): Eigentum in der digitalen Gesellschaft (06.06.2022), Mohr Siebeck, Tübingen, pp. 151-176.

Ostrom, E. (2010): Beyond Markets and States: Polycentric Governance of Complex Economic Systems. *American Economic Review* 100(3): 641–72.

Singh, P. J./ Gurumurthy, A. (2021): Economic Governance of Data Balancing individualist-property approaches with a community rights framework, IT for Change (ed.), http://dx.doi.org/10.2139/ssrn.3873141

Viljoen, S. (2021): A Relational Theory of Data Governance, *The Yale Law Journal*, 131(2): 573-654.

Zygmuntowski, J. J./ Zoboli, L./ Nemitz, P. F. (2021). Embedding European values in data governance: a case for public data commons, *Internet Policy Review*, 10(3). https://doi.org/10.14763/2021.3.1572.

Promises and perils of democratic legitimacy in Social Media Councils

Clara Iglesias Keller
WEIZENBAUM INSTITUE, BERLIN, GERMANY

Theresa Züger
ALEXANDER VON HUMBOLDT INSTITUTE FOR INTERNET AND SOCIETY, BERLIN, GERMANY

This piece approaches Social Media Councils (SMCs) as institutional responses to the democratic deficit of social media platforms. Our first goal is to analyze their promises and perils to advance democratic legitimacy as framed in democratic theory, and thus assess their potential to address the ends that inspired their creation. Against the concentration of power of private platforms (van Dijck, Nieborg, and Poell 2019; Cohen 2019) and depending on their format, SMCs represent an opportunity to improve compliance with procedural guarantees, or to include more stakeholders in platform-led decision-making processes. Meanwhile, their perils lie in the possibility that, instead of indeed promoting the idea of democratic legitimacy that founded them, SMCs end up contributing to validating private self-regulatory institutions that might improve internal procedures, but bring little remedy for existing power imbalances. Against this assessment, our second goal is to outline relevant possibilities for governance designs of SMC and discuss them regarding their potential to create a more legitimate governance of social media platforms. In particular, we will focus on enhancing input legitimacy of such institutional forms, all the while supported by the concept of digital citizenship. Increasing input legitimacy, as we see it, is the necessary first step to increase the legitimacy of platform power overall.

Digital platforms have developed wide reaching political power, as they affect the lives and rights of citizens worldwide. They "play a unique public role in society, shaping culture, economies and societal relationships" (Haggart and Iglesias Keller 2021) and their effective power can be seen as an aspiration to govern in ways which were reserved for states alone in the past. Social media companies are some of the private actors that "aspire to displace more government roles over time, replacing the logic of territorial sovereignty with functional sovereignty", subjecting users to ever more "corporate, rather than democratic, control" (Pasquale 2018). In social media, this control is notably exercised through content moderation, as the set of practices and policies through which platforms interfere in access to information and the right to freedom of expression (a task that encompasses the collection and treatment of personal data). As private companies, they do so with no bound to a democratic process, which has been broadly scrutinized by the literature (Klonick 2020, Cohen 2019, Gillespie 2018).

In recent years, SMCs were increasingly established under the premise to help rectify these deficits. As we argue in the following, they can play an important role in the direction of a stronger democratic legitimation of plattforms's power to address this democratic deficit. However, many of the existing examples still inherit structural shortcomings which leave wide areas of democratically illegitimate power untouched.

Contextual remarks – SMCs as a "piece of a puzzle"

In order to analyze SMCs potential, we start by contextualizing their development in relation to the greater platform governance ecosystem. Overall, platform governance encompasses several "layers of

governance relationships structuring interactions between key parties in today's platform society, including platform companies, users, advertisers, governments, and other political actors" (Gorwa 2019). Due to a "distinct redistribution of governance capacity away from states" (Wagner 2013), platform governance has long been materialized in informal and hybrid forms of rule making. The growing relevance of digital platforms has enhanced claims for embedding their operations with law enforcement and democratic legitimacy standards (Haggart and Iglesias Keller 2021), both through co-regulatory initiatives (Marsden 2011) and state regulation (exemplified in Europe by the recently approved Digital Services Act). Along these initiatives, we also witness the emergence of a hybrid governance landscape formed by institutions that were "developed in a consent based fashion independently, although occasionally steered by state action" (Gorwa, Iglesias Keller, Ganesh 2022). This "hybrid governance" landscape can be seen as a continuation of the important developments in platform governance that have unfolded outside national states' capacities, through voluntary, informal and collaborative governance arrangements (Gorwa 2019, Douek 2019). From expertise focused and industry initiated bodies, like the Global Internet Forum to Counter Terrorism (Ganesh 2021), to governmental or co-regulatory bodies, this landscape includes a variety of institutions marked by different motivations, initiators, members, competences, and levels of state participation (including none).

Against this background, SMCs arise as institutional initiatives distinguished by their aim to improve legitimacy in social media, by providing a solution to companies' "unchecked system" for users' speech governance (Klonick, 2020, p. 2476; Douek, 2019, p. 46). However, even within the very idea of SMCs we can find different institutional forms that operate on different premises and promote different ideas of democratic legitimacy. For instance, platform-initiated SMCs are usually implemented on single companies' own terms, to bolster their internal processes with specific procedural safeguards. We want to call these **internal SMCs**. Meta's Oversight Board, as an example of such, is meant to provide a second instance to Facebook content moderation decisions and to issue recommendations on its policies. Similarly implemented by an internal process, Twitter's Trust and Safety Council gathered a group of organizations that advised on the development of its products, programs and rules. The council was dissolved by Elon Musk after he took charge of Twitter, which exemplifies one of the core problems with internal SMC - their dependance on the companies' leadership decisions. Despite its resolution, we want to recognize the Trust and Safety Council as one of the probed models of an SMC. Twitter's approach differed from Meta's Oversight board, since the Council was constituted by civil society organizations and not staffed by handpicked individuals by the company. However, it was still entirely self-regulatory, which depends (as Elon Musk demonstrated) on the good will of the platform and yet does not effectively change input legitimacy.

In a different approach, which we want to call **external SMCs**, proposals for cross-platform multistakeholder SMCs - like the one from Article 19 - focus on voluntary-compliance mechanisms that provide "a transparent and independent forum to address content moderation issues on social media platforms on the basis of international human rights standards." Despite also being pinned on self-regulation, this version of SMCs presumes "independence from social media companies and participation of representatives of various stakeholders to regulate the practices of the sector" (Article 19).

These (mostly unaccounted) variations of ways through which SMCs can improve legitimacy reflects what is indeed an issue to the approach of democratic legitimacy in platform governance. As previously pointed out by Haggart and Iglesias Keller (2021), there is a need for a multidimensional approach to assessing governance in this field, since focusing on single specific aspects - like multistakeholderism, procedural principles of the rule of law (Suzor 2019) or even international human rights-based

frameworks - reduces the complexity of platform-governance legitimacy, particularly as it relates to democratic legitimacy. Moreover, it overlooks the importance of infiltrating these policy-making processes with safeguards regarding how these decisions are made, and by whom.

Breaking down democratic legitimacy in SMCs

Therefore, we apply the framework tailored by Haggart and Iglesias Keller (2021), according to which democratic legitimacy in the realm of platform governance is not a one-dimensional concept, but rather as a multifaceted phenomenon that must be assessed as such. The authors build on Vivian Schmidt's (2013) modification of Scharpf's (1999) taxonomy of democratic legitimacy as it relates to the European Union - i.e., in an exercise of drawing democratic legitimacy outside of the nation-state. According to this framework, the democratic legitimacy of a policy regime can be divided into three parts. As per Scharf's original contribution, input legitimacy refers to the "responsiveness to citizen concerns as a result of participation by the people," while output legitimacy refers to the "effectiveness of the policy outcomes for the people (Schmidt, 2013, p. 2). To this, Schmidt adds a third category, "throughput legitimacy," which highlights the quality of the governance process and "is judged in terms of the efficacy, accountability and transparency of the (...) governance processes along with their inclusiveness and openness to consultation with the people" (Schmidt, 2013, p. 2).

As a sphere of policy making beyond state's capacities, platform governance legitimacy assessments can profit a lot from Schmidt's framework, especially with regard to the dynamics between the three different strains of legitimacy (Haggart and Iglesias Keller 2021). Although conceptually separate, these different forms of legitimacy interact in ways that may reinforce or undermine each other, and this nuance is mostly overlooked in platform governance policy analysis. For instance, high input legitimacy may be able to compensate for low output legitimacy. Low input or low output legitimacy can mitigate eventual positive effects of high throughput legitimacy. "Importantly, high throughput legitimacy tends to have a minimal effect on input and/or output legitimacy, while low throughput legitimacy tends to have a negative effect on input and/or output legitimacy" (Haggart and Iglesias Keller 2021).

The latter outtake is of particular importance for SMCs, as they largely rely on improving procedures. However, as Schmidt's framework shows, low input and output can compromise the effectiveness of policies that hang on throughput, and this might just be the case for some key SMC conceptions. Adding a second adjudicatory instance to content moderation, increasing transparency or diversifying policy advice has little potential to alter the power status quo if the rules and policies, as well as these throughput legitimacy standards, are still determined by the same players. In this sense, SMCs follow the tendency of other limited platform governance initiatives that "rely excessively on throughput legitimacy for their overall legitimacy", when in fact, they should also "focus on whether they involve rule by and for the people, not just on processes" (Haggart and Iglesias Keller 2021).

Lack of people participation and of public awareness not only decrease the democratic legitimacy of SMCs, but can also compromise whatever developments they allow in the realm of throughput legitimacy. Ultimately, their essence of democratic legitimacy will be more rhetorical than concrete. To avoid this, we further elaborate on ways to improve input legitimacy in SMCs.

Who are "the people" represented by an SMC?

Beyond arguing along the lines of our claim to further input legitimacy in platform governance, we introduce the idea of digital citizenship as a helpful concept. In particular, it addresses the problem that any body of people governed requires an identity and self-understanding, as argued by Schmidt (2013).

For any process of democratic governance, one essential question to begin with is, who are "the people" - both governed and in power? And in the case of platform governance? The idea that platform users are the governed body and that user-engagement can be seen as a participation providing democratic legitimacy is ultimately insufficient. Many of the effects of platform power, like surveillance and a change of communication structures, affect people despite having chosen themselves to use a specific platform. User-engagement can "still not be equated to the expansion of the legitimate power of citizens" (Haggart and Iglesias Keller 2021).

Instead, the idea of the "digital citizen" might be helpful to the question of who "the people", whose input legitimacy should be improved, is in this scenario. The concept of digital citizenship is far from being a universally defined and agreed upon term, but as Jørring Valentim and Porten-Cheé (2018) phrase it: "The concept of digital citizenship has the potential to capture the shifting role of citizens under online conditions." Some understandings of digital citizenship reduce it to the literacy of using the internet ("ones' abilities to access, use, create, and evaluate information and to communicate with others online". A second approach focuses on "the ethics of internet user's behaving appropriately, safely, ethically, and responsibly" (Choi 2016). A third understanding introduces digital citizenship as "different types of online engagement, including political, socio-economic, and cultural participation" (Choi 2016) which most often still refers to state governed politics rather than the self-governance of digital citizens beyond the state.

Last but not least, digital citizenship can also be seen as "another front in citizens' struggle for justice" (Emejulu and McGregor 2019). Understanding citizenship in this context encompasses not only as a set of given rights and duties in relationship with the state but also a claim to new rights and means for participation, as Isin & Ruppert describe in their book "Being digital citizens" (2020). Even though they argue that digital citizenship is a citizenship yet to come, they allow for the possibility of rights claims being made by all humans as potential digital citizens towards old powers of state institutions, as well as new power structures such as tech companies and platforms. This understanding, which is often framed as radical digital citizenship, "is the insistence that citizenship is a process of becoming - that it is an active and reflective state for individual and collective thinking and practice for collective action for the common good. Radical digital citizenship is a fundamentally political practice of understanding the implications of the development and application of technology in our lives" (Emejulu & McGregor 2019). It "is to resist the idea that a neutral technology exists" (Emejulu & McGregor 2019).

As Schmidt envisioned, the governed people would need "a sense of collective identity and/or the formation of a collective political will" (Schmidt, 2013, p. 5). A collective identity of a digital citizenry certainly does not already exist entirely, but we think that it is also an undeniable fact that most humans living today are confronted with a digitized world and oftentimes with a rising awareness that the technology which is the digital infrastructure of this world is not neutral towards them and others. Digitization impacts people's lives very differently but most often widely. Most humans living today act in a digital world in some way. Despite the question, if people are making use of it, in the spirit of human rights they have a right to deliberation, participation and for collective self-governance regarding the digital world. One could say that also the identity of a digital citizenry is a body "yet to come", a dynamic public that is in a process of formation, and constant re-formation.

How can Social Media Councils be designed to start addressing the democratic deficit that platform's power creates?

To answer this question, we outline relevant possibilities for governance designs of SMCs and discuss their potential to create a more legitimate governance of social media platforms. While our focus here is on features that could potentially improve input legitimacy, we have also listed throughput and output related ones, considering how these different fronts interact.

Firstly, the existence and structure of SMCs could be mandated by law, which would give their existence and governance structure a due process, setting out the rules for participation in and communication to such SMC. Moreover, once bound by legislation, SMCs would be within nation-states' democratic design, supported by the flux of legitimacy incoming from popular vote. Such a normative form would be no novelty either for regulatory theory or communication and technology policies, where a number of public-private hybrid arrangements run under the label of co-regulation (Iglesias Keller 2022). Here, Ayres and Braithwaites' idea of "enforced self-regulation" stands out, as the cases where private actors are mandated by government to elaborate on a set of rules aiming at specific purposes (1992).

Secondly, to serve as a legitimate organ, such boards should be constituted in such a way that they ideally represent all stakeholders' interests. The concept of multistakeholderism, which is long practiced in internet governance is helpful in this regard, while the question remains, how to identify and include *all* relevant stakeholders. One way to approach this, could be an inclusive process prior to the actual constitution of the SMC arranging participatory events to identify and recognize local stakeholders.

To meet all digital citizen's needs, an SMC governance structure must be tailored to global and local challenges and interests. This requires a set-up on the national level. International governance structures might be relevant but we will leave them out of the scope of this paper for it exceeds the level of complexity we can address here. If the goal is to make SMC responsive to citizens' concerns as a result of participation, channels must exist that allow for citizens' concerns to be heard and for threats to human rights to be considered.

The national board could be a mix of

- political representatives, nominated by the executive power and approved by Parliament, as it is the case in several independent specialized regulatory bodies;
- a number of elected citizens (users and non-users) of the platform, meeting broad diversity to increase a sense engagement for citizens and allow for more direct representation;
- a number of local advocates of vulnerable groups; and
- potentially changing guests to the board processes, e.g. academics, journalists, activists or human-rights advocates to represent in specific issues.
- Additionally, the framework of the cross-platform SMC would need to be designed in a way to allow for equal terms of participation.
- For this, all board members must be equally paid for their efforts and reimbursed for missing their normal jobs, in the case of temporary participation; legal oversight needs to be organized to ensure due process of the SMC election, and
- Continuous work of the council membership in the board could be restricted to the given election period of the parliament to ensure the change of personnel.

Lastly, what rights and setup should the board have?

- The SMC should have far reaching rights, which are not necessarily restricted to internal functions of the platform such as content moderation but also strategic decisions of the platform policies that might affect citizens;
- the SMCs work could be organized as interventional acts, meaning they have the right to question or veto a decision in the process and intervene in different types of decisions of the SM platform on a national level;
- Regarding content moderation, their competences would ideally include not only oversight over second-instance to content moderation decisions, but also the possibility of reviewing the internal policies upon which these decisions are taken (including terms of use, community standards and their update);
- Moreover, SMCs could be equipped to receive and process complaints brought to their attention by users or non-users of the platform through different, easy to access channels. This could include channels for dialogue between the public and the SMC, to be established on a periodic basis or not;
- Decisions of the SMCs should (with possible exceptions for privacy reasons) be documented and made public, to increase public awareness of the SMC;
- Also, to increase public knowledge and transparency of the SMC, council meetings should be open for visitation by the public at certain hours and in appropriate settings.

As it shows, such a design of an SMC is meant to serve the specific purpose of promoting a more comprehensive idea of democratic legitimacy of social media governance. Depending on their specific competences, SMCs potential to fulfill this goal will vary, but should not be expected to rectify all democratic shortcomings of social media. Our approach is not to be taken as the single way for social media platforms to become spaces driven by the public interest, but a proposal to address more severe democratic deficits and improve institutional forms currently implemented in such spaces.

References

Article 19, Social Media Councils. Available at: https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf. Accessed on: 01 dec. 2022.

Choi M. (2016). A concept analysis of digital citizenship. Theory & Research in Social

Education, 00:1–43, DOI: 10.1080/00933104.2016.1210549

Cohen, J. E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190246693.001.0001

Emejulu A. & McGregor C. (2019). Towards radical digital citizenship in digital education.

Critical Studies in Education, 60:1, 131-147, DOI: 10.1080/17508487.2016.1234494

Ganesh, B. (2021, March 16). *Platform Racism: How Minimizing Racism Privileges Far Right Extremism* [Social Sciences Research Council]. https://items.ssrc.org/extremism-online/platform-racism-how-minimizing-racism-privileges-far-right-extremism/

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Iglesias Keller, C. (2022). The perks of co-regulation: An institutional arrangement for social media regulation? In E. Celeste, A. Heldt, & C. Iglesias Keller, *Constitutionalising Social Media*. Hart Publishers.

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, 131, 1598–1669.

Marsden, C. T. (2011). *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace.* Cambridge University Press.

Oversight Board, Ensuring respect for free expression, through independent judgment. Available at: https://www.oversightboard.com/. Accessed on: 01 dec. 2022.

Pasquale F. (2018). Frank Pasquale on the Shift from Territorial to Functional Sovereignty, https://blog.p2pfoundation.net/frank-pasquale-on-the-shift-from-territorial-to-functional-sovereignty/2018/01/16

Scharpf, F. (1999). *Governing in Europe: Effective and Democratic?* Oxford University Press.

Schmidt, V. (2013). Democracy and legitimacy in the European Union revisited: Input, output and "throughput", *Political Studies, 61(1)(2013), pp. 2-22.*

Wagner, B. (2013). Governing Internet Expression: How Public and Private Regulation Shape Expression Governance. *Journal of Information Technology & Politics*, *10*(4), 389–403. https://doi.org/10.1080/19331681.2013.799051

Social Media Councils under the DSA: a path to individual error correction at scale?

Aleksandra Kuczerawy*
CENTER FOR IT AND IP, KU LEUVEN, BELGIUM

Introduction

Content moderation by online social media is an arduous process. It consists of private entities making decisions that impact the effective exercise of the right to freedom of expression, oftentimes without meaningful accountability. It also tends to upset everyone who, even remotely, cares about the outcome. Social Media Councils (SMCs) have been presented as a solution to resolve this conundrum. The idea has not been widely picked up so far, although examples can be found. In Germany, a self-regulatory NGO "Freiwillige Selbstkontrolle Multimedia-Diensteanbieter" (FSM) has been set up by YouTube and Facebook, under the NetzDG, to decide difficult content removal cases. The most prominent example, of course, is the Meta Oversight Board. Both bodies resolve only a small selection of cases and do not fully reflect the original idea of SMCs. But is this about to change thanks to the Digital Services Act (DSA)? In an attempt to regulate content moderation practices of online platforms, the DSA mandates the use of internal complaint-handling mechanisms as well as external out-of-court dispute settlements for individual error correction. This essay looks at both paths to remedy to see if they align with the original idea of SMC. Could SMCs fit the framework of the DSA? This essay will focus on two aspects in particular, namely independence and scalability. Furthermore, this essay looks at the special case of the Meta Oversight Board (MOB) to answer whether, and under what circumstances, the MOB could serve as an out-of-court dispute settlement body within the meaning of the DSA.

Social Media Councils in a nutshell

In 2018, the UN Special Rapporteur for Freedom of Speech David Kaye recommended that "all segments of the ICT sector that moderate content or act as gatekeepers should make the development of industry-wide accountability mechanisms (such as a social media council) a top priority". It was a strong endorsement of Social Media Councils (SMCs), an idea that had been propagated by Article 19 to address the unaccountable power of social media over people's right to freedom of expression. According to the original concept, a Social Media Council is a multi-stakeholder voluntary-compliance mechanism for the oversight of content moderation practices on social media. An SMC would provide a transparent and independent forum to address content moderation issues on the basis of international human rights standards. An SMC would have several roles. For starters, it would serve as an appeal body for individual content moderation decisions. It would also provide general guidance on content moderation policies in line with international human rights standards and act as a forum where all stakeholders could discuss and adopt recommendations. Compliance with SMC decisions and recommendations would be on a voluntary basis. Its core characteristics include independence from government and from any particular social media company. It should include representatives from all

relevant stakeholders, such as media associations, media regulatory bodies, freedom of expression experts, academia, and civil society appointed through a democratic and transparent process. Moreover, SMCs should receive a stable and appropriate level of funding to ensure its independence and capacity to operate. Funding could come from social media companies (at least partially), as well as public subsidies, other stakeholders, or philanthropic organisations. Ideally, SMCs would operate on national (or regional) basis to ensure participation of decision-makers with the best knowledge of the local context and understanding of cultural, linguistic, historical, political, and social nuances.

So far, the idea has not been widely picked up, even though it received considerable praise from academics, international policy experts, and civil society organisations. Is it, perhaps, about to change with the arrival of the DSA? Can the provisions on redress mechanisms encourage the implementation of the SMC and facilitate its operationalization?

Remedies under the DSA

Interestingly, during the negotiation process of the DSA the idea of establishing Social Media Councils was contemplated. A group of MEPs in the IMCO committee proposed an amendment (1686) that would require the establishment of an independent advisory group, representing recipients of the service, or groups potentially impacted by services. The European Social Media Council (ESMC) would be independent of commercial interests, with expertise and competence in the area of international human rights law, content moderation, algorithmic systems, media, consumer protection, disinformation, hateful speech, etc. The ESMC would issue non-binding guiding principles and recommendations to improve content moderation processes; foster a participative and transparent public debate around content moderation; and issue policy and enforcement recommendations. The ESMC, therefore, would not carry out any review or correction of individual decisions. The proposed amendment, however, did not make it into the final position of the European Parliament.

The DSA, aiming to strengthen fundamental rights online, focuses strongly on the right to freedom of expression but also the right to effective remedy. This latter right, in the context of content moderation, is translated into a requirement that any erroneous decision regarding content should allow for rectification. The DSA offers three different redress mechanisms for content moderation decisions: internal complaint-handling mechanisms (Art. 20), out-of-court dispute settlement (Art. 21) and judicial redress (no specific provision but several mentions). They can be used in sequence or separately, as self-standing mechanisms.

According to Art. 20, providers of online platforms should create an internal complaint-handling system and make it available for at least six months from the time a measure against a piece of content was taken. The complaint-handling system should be easy to access and user-friendly. It should handle complaints in a timely, non-discriminatory, diligent and non-arbitrary manner. And most importantly, review by such a complaint-handling system should lead to a reversal of a previous content moderation decisions, if the complaint contained sufficient grounds to justify such reversal.

For disputes where a platform stands firmly behind its action (e.g. because it is in line with their internal policy), the DSA foresees an external redress mechanism. For that purpose, Art. 21 introduces the out-of-court dispute settlement. It is meant for complaints that could not be satisfactorily resolved via the internal mechanism, or would not be resolved via the internal mechanism because the complainant went straight for this second option.

A complainant should be able to select any out-of-court dispute settlement body certified by a competent Digital Services Coordinator. Such certification is then valid in all EU countries. Conditions for

certification include impartiality, and independence, including financial independence and independence at the level of the persons resolving disputes. Certification also requires the necessary expertise in one or more areas of illegal content or application of terms and conditions. The rules of procedure under which the settlement body would operate should be clear, fair and easily accessible. The bodies should reach a decision within a specified time frame (90 days with possible extension of another 90 days for complex disputes). The costs of the procedure are spread between the platform and a complainant, depending on whose claim succeeds. But the main financial burden is on the platforms, while the users should be able to access the settlement bodies free of charge or at a nominal fee.

Parties to the dispute should engage in good faith with the selected body in an attempt to find a solution. The dispute settlement bodies, however, do not have the power to impose a binding settlement of the dispute on the parties.

Match or no match?

So how do these two redress mechanisms correspond with the original idea of SMCs? And more specifically, can they constitute a pathway to operationalize the SMC idea in fulfilling its intended role in individual error correction?

Internal complaint handling systems are, arguably, a necessary first instance redress path that offers a possibility of a quick dispute resolution free of charge. It is also, usually, able to handle a large amount of complaints. The main focus of the internal review is a compliance assessment with the platform's existing policies, rather than questioning if the policies are in fact reasonable. Although important to efficiently resolve (at least some) content moderation disputes, this mechanism clearly does not amount to an SMC. Internal complaint mechanisms are, as is explicitly stated, internal. This means that they are administered by the platform, financed by the platform, accountable to the platform and dependent on the platform's choices regarding content policies. It is therefore, everything but independent from the platform.

The out-of-court dispute settlement, on the other hand, seems to be a better match. First of all, the out of court settlement bodies under the DSA would be external, outside of the platform's own structure. Like for any other ADR mechanism, this is a crucial element and it aligns the requirements for an SMC. The DSA does not make any demands regarding the composition of the settlement body. It might be difficult for any such body to achieve the ideal composition of an SMC that calls for broad participation of all relevant stakeholders (i.e. media associations, media regulatory bodies, freedom of expression experts, academia, and civil society appointed through a democratic and transparent process). But it could definitely strive for a broad representation of interests. In any case, as long as the decision-making members are independent and impartial, the body would still be able to fulfill its intended role.

Another similarity between SMCs and dispute settlement bodies under the DSA is that the decisions of the SMC and the settlement bodies shall not be binding. The idea of an SMC is based on the tradition of self-regulation and voluntary participation. No legal obligations should be necessary, therefore, to ensure voluntary compliance with the SMC decisions. Arguments for not making the decisions binding under the DSA were discussed during the negotiation process. It was argued that binding decisions would create extra-judicial bodies, would lead to must-carry orders and would prevent platforms from obtaining redress. Either way, the non-binding approach corresponds with the general requirements articulated in Directive 2013/11 on alternative dispute resolutions.

The DSA requires certification of the settlement body by (at least one country's) Digital Services Coordinator and specification of the languages in which the procedure is available. In theory, that could

mean only one dispute settlement body for the whole EU. Considering the amount of languages and cultures, as well as types of content to cover, a variety of specialized bodies (thematically, geographically and linguistically) would be more likely. This d the requirement that the decision-making in SMCs happens at a level that would ensure knowledge of local context and nuances.

According to the original concept, a complaint to the SMC would be possible only after all possibilities of remedying the issue with the social media company have been exhausted. In contrast, a complaint to an out of court settlement body can be initiated at any stage, even without using the platform's internal complaint mechanism first. The difference in approach here, however, is not a source of major incompatibility and mainly impacts the number of potential complaints. Charging users for using the out of court dispute settlement procedure, even if at nominal fee, could have a similar limiting result (especially since the internal mechanism can be used first, for free).

Effectively, the out of court settlement bodies whose role would be to review specific content moderation decisions, upon complaint by those affected by the decision, closely resemble the original SMC concept. This is also a view held by the proponent of the original SMC idea, Article 19, which indicated that a SMC could serve as the out-of-court dispute settlement mechanism required under the DSA.

It should be highlighted, however, that a dispute settlement body under the DSA would fulfil only one of the roles foreseen for SMCs, namely the rectification of errors in individual instances. It would not function as a body designed to provide guidance or give recommendations to social media platforms.

Independence is key

One element that keeps coming back in the description of both SMCs and dispute resolution mechanisms is the requirement of independence. But which facets of independence should be considered in the context of content moderation disputes? And since we're talking about private enforcement mechanisms, why does it even matter?

Access to an independent and impartial tribunal is at the core of the right to fair trial and the right to effective remedy. Coming from international human rights instruments, such as the ECHR and CFREU, these rights are primarily addressed to States, instructing them how to organize their judicial system. In the context of private enforcement, these rights serve mainly as guideposts. An ideal to look up to, without an obligation of strict compliance. Yet they should be given careful consideration, as the aim of the out of court settlement mechanisms is, in fact, to provide effective remedies.

To be clear, neither the out of court dispute settlement bodies under the DSA, nor the SMCs can be considered independent tribunals in the understanding of art. 6 ECHR. Apart from being private bodies, they do not have the power to issue binding decisions. Nevertheless, it is still worth looking at how the ECHR and ECtHR understand independence. It is also worth checking how this concept has been translated for the purpose of assessing the independence of regulators.

First and foremost, independence is understood as a guarantee that there is no undue influence on the decision-making process. It is a condition for parties to accept the outcome, trusting that the process was fair and there was no external interference. Independence refers to freedom from interference by both the executive power and by the parties that may have interests in the outcome. Independence could be impacted in different ways. It is possible, therefore, to distinguish components, such as the manner of appointment of the members of the decision-making body, the existence of sufficient safeguards against the risk of outside pressures, as well as financial freedom. There is no formula to calculate independence

that specifies the weight of individual components. Depending on the circumstances, deficiencies in one area (e.g. financing by a party subject to the decisions) may be compensated by other elements (e.g. existence of strong safeguards against undue influence).

One component of independence that can be hard to assess is the appearance of independence. Appearance of independence comes from the idea that "justice must not only be done, it must also be seen to be done". The main purpose is maintaining confidence and trust of the public in the independence of the decision making body. The appearance of independence may be compromised when one of the parties has reason to believe that independence of the decision-maker is not guaranteed and it may be triggered by various factors. This fear, however, must be objectively justified.

In the context of private enforcement, independence is crucial to legitimize operations of such out of court dispute settlement bodies. The source of financing as well as the manner of appointment of the members must therefore be clear and transparent. Strong involvement of one party in either of those elements does not necessarily exclude independent decision-making but it must be accompanied by sufficient safeguards against the outside pressure. It might impact, however, the appearance of independence. A complainant might question, for example, the independence of a body that happens to be fully financed by the opponent in the conflict. External positioning of the out of court dispute settlement bodies alone does not always guarantee independence but it strongly supports it. Adding additional degrees of separation between the decision-makers and those affected its decisions contributes to the appearance of independence. It also helps to increase confidence in objectivity and impartiality of decisions. Providing clear, fair and accessible rules of the procedure ensuring that the process of resolving conflicts is handled in a fair manner is another element that improves the trust and, as a result, perception of legitimacy.

Meta Oversight Board as a global cross-platform dispute settlement body?

The remaining question is whether the Meta Oversight Board (MOB) could become a body that satisfies the requirements of an SMC and a dispute settlement body under the DSA.

Dubbed as Meta's own Supreme Court, the MOB is an example of voluntary corporate self-regulation. Its role is to review a selection of Facebook's and Instagram's removal decisions, to oversee the enforcement process and to provide policy recommendations. The cases handled by the MOB are selected by a committee that focuses on cases that raise important policy questions across the platform, that have a major effect on public discourse, and have impact on large numbers of users. The number of cases that the MOB has actually analyzed remains limited. To illustrate, in Q2 of 2022, out of 347,304 cases submitted the MOB has resolved three.

In order to ensure the independence of the MOB's decision-making process, several structural and procedural measures have been introduced. While the first members of the MOB have been selected by Meta, the MOB will select its future members on its own. The members are not Meta employees and cannot be removed by Meta. The initial members of the MOB include a Nobel Prize-winning press freedom activist, a former prime minister, and several legal academics and former judges. Funding of the MOB comes from Meta, but it has been organised through a trust to protect the independence. Moreover, the MOB operates according to a charter and bylaws, describing its operational procedures. With all these precautions, and despite criticism of the whole concept, the independence of the decision-making process of the MOB is generally not questioned.

Despite all these properties, the MOB is not actually an SMC within the original understanding of the concept. The initial excitement over this endeavor has somewhat faded, and it is currently perceived more of a PR move created to legitimize Meta's content moderation practices. But could the MOB become something more and provide an answer to both civil society and EU policymakers calls for independent oversight? In particular, could the MOB serve as an out-of-court dispute settlement body within the meaning of the DSA, also for other online platforms?

In theory, the MOB could evolve to fulfill the role of a dispute settlement body within the meaning of the DSA. The number of changes it would have to undergo, however, are rather significant. First of all, Meta would need to make even more concessions to distance itself sufficiently from the new body to maintain the appearance of independence. Moreover, keeping the focus only on Meta's activities would be incompatible with the requirements listed in the DSA.

Moreover, the MOB would have to become certified and established in at least one EU country. Ideally, it should be operating in multiple EU countries to make sure that all the languages as well as specific social, cultural, historical and political nuances can be properly taken into account. In addition, the process of selecting the members and the financing structure would have to change, especially if this new version of the MOB was to review content moderation decisions not only by Facebook and Instagram but also other social media platforms. And here comes a major difficulty that the new MOB would encounter: the number of decisions to make. From a body that very selectively picks a handful of cases to review, it would have to become a body that is able to review each individual complaint. And it would need to be able to decide not only complaints about removals and non-removals but also complaints about any other action taken against content, such as restrictions on visibility, suspension or termination of a service or an account, and decisions restricting the ability to monetize content.

The amount of people required to cover the incoming complaints would be enormous. For this purpose, the current MOB structure cannot be replicated at scale. Workforce appropriate in size to process all the complaints cannot be filled with Noble-prize winners and law professors receiving 6-figure salaries. After all, it is not the idea of the DSA to end up with another layer of low-paid overworked content moderators processing complaints against the clock. Unless its operations remain local (e.g. focusing only on one country and one language), creating an out of court dispute settlement body of this size would basically require Meta to open a completely new and parallel business venture. While it is not impossible, of course, the necessary changes would be far from trivial. And the final result would differ significantly from the MOB in its current form.

Conclusion

The idea of SMCs is still waiting to be picked up and implemented in practice. With the help of the DSA, it could be operationalized to become a more realistic tool able to address the problem of unaccountable decision making with an impact on fundamental human rights. Even though there are some differences between the original SMC concept and the out of court dispute settlement mechanism under the DSA, they are not fundamentally incompatible. The main challenge would be to earn the trust and confidence of the public that such a body is able to review the content moderation decisions in a fair, independent, and efficient manner. Only then, the public would refer their grievances in a search of effective remedy for content moderation decisions. The MOB, at the moment, only vaguely resembles the original SMC idea. In order to become an SMC that would also fulfill the requirements of Art. 21 DSA, it would have to undergo a massive makeover. For starters, it would have to significantly grow in size to become a mechanism for individual error correction. It would also have to distance itself even

further from Meta to maintain the appearance of independence. The necessary changes would make it nothing like it is right now. But perhaps for the better?

Social Media Councils as Self-Regulatory Bodies

Riku Neuvonen
TAMPERE UNIVERSITY, UNIVERSITY OF HELSINKI, FINLAND

Introduction

In principle, Social Media Councils are self-regulatory bodies; in this context, self-regulation can be defined as regulation of the private sector by nongovernmental entities. Hence, it is the opposite of public regulation, that is, hard regulation (Black 1996). During the past few decades, these modes of regulation have been mixed, with the hybrid regulatory approaches being termed coregulation. Coregulation has been typical for audio-visual media, in which classification systems have been enacted by law or at least endorsed by authorities (Marsden 2011, 157–160). In the EU, innovative regulation has been part of the so-called 'new approach' towards legislation, in which industry-led standardisation and regulation are the key components. However, regulating public debate is very different from regulating a cucumber trade (Quintel and Ulrich 2020).

In addition to the privatisation of public regulation, private companies and organisations have developed different models of self-regulation. One incentive driving these ventures is the global nature of the internet, which means a lack of standards and overlapping regulations. In fields related to human rights, especially freedom of speech, demands for private bodies to respect and even enforce human rights have grown. The obligation to actively protect rights and the horizontal effect of human rights are disputable, but treaties such as the United Nations Guiding Principles on Business and Human Rights (UNGP) have increased the pressure on companies to do something. The Oversight Board has been Meta's answer to several scandals and regulatory pressures. On the other hand, NGOs such as Article 19 have developed models for Social Media Councils, and there have been several attempts to bring democratic values and processes to the internet.

In the present paper, my hypothesis is that, compared with existing models of self-regulation and coregulation, Social Media Councils can learn from previous experiments but still face challenges that are unique to the environment in which these councils operate. I first demonstrate why there are several self-regulatory and coregulatory bodies in the field of media. My question is why there is a need for self-regulation, here in the form of Social Media Councils. My second question is how other self-regulatory bodies are formed and what legal basis they operate within. The third question relates to what kind of needs the Social Media Councils could address.

Why Self-Regulation?

Self-regulation is often caused by the fear of public regulation. For example, in April 1916, the Swedish newspaper Ny Dagligt published a private letter, causing a scandal in that country. The scandal led to calls for greater regulation of the press. As a result, the first modern press council, Pressens Opinionsnämnd (PON), was established in Sweden in 1916. However, PON only covered the press and suffered from a lack of funding in its early years (Weibull and Börjesson 1995). In Great Britain in 1949, the first Royal Commission on the Press recommended that a General Council of the Press should be formed to govern the behaviour of print media publishers (Frost 2000). The commission was founded amid public concern that a concentration of ownership was inhibiting free expression, leading to factual inaccuracies and allowing advertisers to influence editorial content. In Finland, the first guidelines for

journalists were drawn up in 1958 by Finnish media associations; this effort did not prevent the press from publishing scandalous news. The public blamed a few scandalous publications for causing the death of famous writers. Hence, the Council of Massa Media was established out of fear of government regulation in 1968, but it did not prevent the criminalisation of the dissemination of information violating personal privacy in 1973 (Neuvonen 2005).

Similar developments have been seen with the game industry. Violence in video games caused moral panic in the USA in the early 1990s. As a result, members of the combined United States Senate Committees on Governmental Affairs and the Judiciary held congressional hearings on the subject. The game industry was given options to organise self-regulation or be the object of public regulation. During these hearings, the industry developed a model of self-regulation, the Entertainment Software Rating Board (ESBR). In Brown v. Entertainment Merchants Association, 564 U.S. 768 (2011) the ruling was that law prohibiting the sale or rental of violent video games to minors violated the First Amendment. However, the court stated that the ESBR provides parents with sufficient information about the content of the games, so there is no need for hard regulation (Wuller 2013). The system is similar to other private rating systems in the USA: MPAA/MPA (Motion Picture Association), which is controlled by CARA (Classification and Rating Administration); the RIAA (Recording Industry Association of America) record rating system; and television companies' TV Parental Guidelines.

In Europe, long-running negotiations led to the establishment of Pan European Game Information (PEGI) (Marsden 2011). Before this, games had been regulated and even banned based on film regulations and criminal law in different countries. The EU Commission has repeatedly considered PEGI's activities to promote the goals of human rights, especially a safe environment for children. The classification system has been strongly involved in the recommendations of projects promoting safer internet and media literacy. In its resolution, the European Parliament has stated that, in addition to their disadvantages, games also have clear benefits and are part of the current digital operating environment. Parliament urges member states to continue and intensify cooperation for the development of the PEGI system. According to European Parliament, modes of self-regulation should be sought in the gaming industry, thus avoiding the need for EU-level legislation. The Council of Europe, which is behind the European human rights system, in cooperation with the Interactive Software Federation of Europe, has created the principles for considering human rights in online games. Therefore, it can be said that self-regulation has not only been approved by the EU, but also by the CoE.

Social media has caused several uproars and scandals in the late 2010s and 2020s. Consequently, demands for the regulation of social media have grown. Meta's (Facebook and Instagram) answer has been the Oversight Board, while other platform companies have taken different initiatives to make moderation more transparent and even democratic. The incentives for these initiatives are regulatory demands on both sides of the Atlantic. However, in the EU, the Digital Services Act is now nationally implemented, and it has elements of coregulation such as trusted flaggers and dispute settlement mechanisms.

These developments seem to be similar. The abolition of censorship and the rise of the press as an industry caused the need for press ethics and self-regulation as regulation instead of hard regulation. The significance of game industry growth during the 1990s, but games did not fit into the framework of electronic communication regulation. As a result, games were censored (Germany) or self-regulated (Netherlands), or the regulation authorities did not recognise them for a long time (UK, Finland).

In the early 1990s, the internet was seen as an area of freedom and a good business opportunity. The lack of effective regional or global norms and exemptions of liability, that is, "Good Samaritan" principles, led both sides of the Atlantic to enable the free growth of social media. Today, the empires

are striking back, and now, companies and NGOs are looking for self-regulation as a way to fix the biggest problems and avoid harsher regulations. It should be remembered that self-regulation is not always the best answer. In the US, the Production Code (i.e., Hays Code) for movies and Comic Code were self-regulatory systems but implemented harsher censorship than similar activities by public authorities in different countries.

Who Are the Regulators and Who Are the Subjects of Regulation?

It is necessary to be aware of who the regulators are; especially in self-regulation, there is (usually) no standards or authorisation. Self-regulatory bodies claim that they represent the field in question, but this demands critical evaluation. Another essential aspect is the scope of regulation. The Alliance of Independent Press Councils of Europe (AIPCE) has described press (media) council functions as monitoring the codes of ethics/practice and defending freedom of the press. The AIPCE is the most important union for press councils. The World Association of Press Councils (WAPC) describes itself as a defender of free speech, but none of the members of the AIPCE members are members of the WAPC; instead, members include, for example, Turkey, Kenya, Zimbabwe and other councils from countries not known for free media. In this paper, I focus on the AIPCE.

According to the comparative data on media councils collected by the AIPCE, media councils differ from each other. There is no clear blueprint on how to organise media councils and on what basis. The data from the AIPCE cover 32 media councils. First, four councils have been established by a decree, and two are recognised in law. The Danish Press Council is even a public entity, though it is an independent tribunal. Second, membership in press councils varies. Some have accepted individual journalists as members, and others have accepted only organisations, media outlets or both. Six councils also accept other members than those mentioned before. Third, print newspapers and magazines, as well as websites (of media outlets), are within scope in all of them, but 75% also cover television and radio. However, the mandate of most councils has covered all media only in recent decades.

Management of the PEGI system was handed to PEGI s.a., an independent, not-for-profit company with a social purpose established under Belgian law. The Netherlands Institute for the Classification of Audio-Visual Media and the VSC Rating Board (British) administrate the system on behalf of PEGI. In the PEGI council, 35 media authorities from the member states are represented. In some countries, for example, in Finland, PEGI ratings are accepted in law as legal ratings. Therefore, even though the PEGI is an independent company, it exercises public power and is endorsed and acknowledged by the EU and member states.

Meta set up an irrevocable trust, which is a legal entity in Anglo-American law but not so much in continental Europe. The trust created a limited liability company that owns, facilities and hires staff for the Oversight Board (OB). As Lorenzo Gradoni demonstrated, the OB itself is not a legal entity, and the members of the OB are in a contractual relationship with the background organisation (Gradoni 2022). As an interesting detail, the Oversight Board Trust Agreement allows for the accession of other platforms with the consent of Meta, as long as they are US persons and contribute to the funding of OB. The nature of SMCs is important because SMCs need to acquire facilities and potentially hire staff and pay for expert services to at least cover the expenses of the members. This is important because the (legal) form affects how practical matters are organised, such as accounting, transparency, employee rights and liability.

The proposals for SMCs include reviewing individual content moderation decisions made by social media platforms based on international standards on freedom of expression and other fundamental rights

but without creating legal obligations. This is a similar feature in most media councils. The SMCs should be established via a fully inclusive and transparent process that includes broad and inclusive representation, and SMCs must be transparent. There are many demands that SMCs should be inclusive and democratic, which increases the demands for both the nature of background organisation and for day-to-day management.

When compared with media councils, I have earlier demonstrated that one of the key issues is trust between stakeholders (Neuvonen 2022). It is necessary to define and commit stakeholders and to assure that SMCs benefit all parties. This requires that the initiatives for SMCs define stakeholders with care. Article 19 suggests that the SMC should be made up of representatives from social media companies, media, journalists, media regulators, press councils, the advertising industry, civil society organisations and academics. These groups have been included in the Irish SMC. They might leave open questions regarding inclusivity, third parties and the public. For example, in the Finnish Mass Media Council, eight members represent the media and eight the general public (some of them are members of academia). The Finnish Council is respected and operated for more than 50 years. This means that the SMCs must balance between the requirements of adding all stakeholders, inclusivity and democracy in both background organisation and management without compromising efficiency. It is not possible to fulfil all wishes.

The role of the media councils is to supervise media ethics. These ethical guidelines are connected to the ethics of journalism and journalism as a profession. Journalists and media outlets identify with the ideals of journalism and free speech. An individual journalist has the desire to follow the ethical code, and here, journalists are important stakeholders of self-regulation. Five councils are responsible for the distribution processes for press cards for journalists. According to many studies, moderators are outsourced employees and artificial intelligence is used a lot. There is no clear profession to identify with or professional ethos for moderators. The subject of OB is moderation as a whole and single processes and practices, not for a single moderator as such. An individual journalist is committed to their work, but there is a big gap between the moderator who made the decision and the complaint process. In the proposals for SMCs, moderators are never mentioned as stakeholders, which is understandable because moderation is outsourced and AI handles most of the job. Nevertheless, moderators are often ignored when social media practices are discussed. This also makes a difference compared with media councils.

Another issue is the scope. Most of the media councils cover all media (print, radio, television) but how to define social media. The scope of OB covers Facebook and Instagram, and PEGI requires recognition from states. The concept of platforms also covers food delivery companies, Uber and streaming services. Streaming services could also establish chats, for example. What is the status of chats and forums in games? To be effective, the SMC should cover most social media, but the field is very large, from small discussion forums to social media giants. It should also be noted that there might be competing councils. In the UK, the Leveson Report led to a system based on the 2013 Royal Charter on self-regulation of the press (hereinafter the Charter). The Charter created the Press Recognition Panel (PRP) to ensure that regulators of the press and other news publishers would be independent, properly funded and able to protect the public. The first regulator recognised by the PRP was the Independent Monitor for the Press (IMPRESS). However, none of the large national publishers were members of IMPRESS. Instead, most national publications were members of the Independent Press Standards Organisation (IPSO), which the PRB did not recognise. In addition, several prestigious newspapers (e.g., The Guardian, Financial Times) have established their own independent complaint systems.

What Can Added Value SMCs Create?

Social media has benefitted from the ethos of freedom found in the early days of the internet. However, China and Russia have developed mechanisms to monitor and block internet routes. Similarly, the biggest democracy, India, among other countries, uses shutdowns to control network activities. In Europe, the EU's digital package is changing the rules, and many hope that in a similar way to the GDPR, these regulations can create the Brussels Effect, that is, making regional rules de facto global rules. In the USA, Texas and Florida have introduced social media laws. These laws are already contested in courts, and their nature is very political.

Communications rights, digital rights, epistemic rights and digital constitutionalism, among other buzzwords, are attempts from civil society and academia to create principles and normativity on the internet. SMCs have similar backgrounds, but as self-regulatory bodies, SMCs are more than just frameworks. For example, the OB is a sui generis attempt to create some kind of higher body to handle the fundamental issues of single company platforms. The OB has been compared with the US Supreme Court, especially regarding the Marbury v. Madison case. The fact it can be compared with the Supreme Court is noteworthy because it is also necessary to note the language used in discussions about SMC.

Using legal metaphors and adding the digital into rights and constitutionalism is tempting. However, in the case of SMCs, this could mean moving from self-regulation to the domain of law. Even though the Danish Press Council is a public entity, it maintains a wall between law and self-regulation. In none of the media councils has the government been involved in the appeal process. Media councils are committed to self-regulation, which is seen as separate from the legal system. Indeed, the self-regulation of media is about ethics.

In addition, because of the AIPCE, only four media councils can sanction the media or journalist for an upheld complaint, and only two councils can order financial consequences for the breach of a journalistic principle. It is also necessary to note what grounds complaints can be made upon; for example, in Finland, complaints are quite open, and anyone can complain. In comparison, in Sweden, the grounds for complaints are more regulated, including the eligibility as a complainer.

The references to Marbury v. Madison can be seen as a part of digital constitutionalism. Digital constitutionalism can be divided into two sects: the internet's own (techno-utopian) legal system or the growing importance of existing constitutional systems on the internet. The proposals for SMCs highlight the importance of human rights, especially UN human rights. This means that SMCs can be seen as human rights interpreters. The OB refers to the UNGP and the International Covenant on Civil and Political Rights (ICCPR). Here, the UNGP requires companies to follow and respect human rights, especially regarding the ICCPR. However, this does not constitute a horizontal effect of human rights. Therefore, using legal metaphors and the language of law, SMCs are considered more quasi-courts than self-regulatory bodies. There are many nonlegal and private human rights actors, so what can new SMCs bring in? Or should SMCs have guidelines of their own?

There are several trends that have been occurring. First, the internet can be seen as a system of its own, and since Barlow's declaration, many have thought that the internet community itself could achieve sovereignty. Second, states and supranational entities are controlling the internet more intensively. Third, traditional media councils are increasingly engaged with social media, and media regulation is increasingly affecting media councils. Where is there room for SMCs? Should SMCs remain in the private world of self-regulation, or is coregulation a more suitable form?

Finally, there is the question of global versus local. Here, I will make a distinction that is more specific to global, regional (i.e., Europe), local (state) and hyperlocal (community). The PEGI is a regional

European operator and requires recognition from the member states. Media councils are national, and some councils are hyperlocal; for example, in Belgium, there are the Council for Journalism in Flanders and the Council for Ethical Journalism in Wallonia or the Information Council of Catalonia. The Swiss Press Council members represent different language regions. Looking at these examples, we need global, at least regional, rules for local and hyperlocal problems (Azzi 2021). According to Facebook Files, Facebook and Instagram have focused their moderation resources mostly on the US and English-speaking world. This can lead to problems. For example, The Guardian revealed during the COVID-19 pandemic that there has been a lot of misinformation and fake news in Spanish, which has the second most native speakers in the world. How many resources are there for Finnish, Latvian, Basque or Sami speakers is a mystery. How can SMCs strike a balance between global and regional impact and regional and linguistic representativeness? Without representativeness, there is no commitment of users.

Conclusion

In Ireland, SMCs have been interesting experiments both intellectually and in practice. I have compared the proposals and ideals for SMCs with existing self-regulatory and coregulatory bodies. My conclusions are based on the idea that Social Media Councils would be established. SMCs can be different, and there can be several of them.

1) The need for SMCs is similar to the pressure to establish media councils and organise the classification of games other than through hard regulation. Both platform companies and civil society are looking for solutions for content-monitoring issues in social media. However, at the same time, the EU and other legislative bodies are tightening the regulation of social media.

2) It is important to note the legal nature of the SMC or its background organisation. The Oversight Board does not exist in normal legal standards, but the company and trust operate its day-to-day management and funding. The OB itself operates in a void called global law. As a legal person, the SMC must be established in some state. This will affect how to make contracts, requirements for audits and the status of employees.

3) The SMC's credibility requires representativeness. The most effective media councils have representatives from different groups. All stakeholders must be engaged, and the public must be represented. However, how do we account for moderators?

4) It is necessary to clearly think of who can make a complaint and on what grounds.

5) Should we think that SMCs are courts? The media councils have drawn a clear line between self-regulation and law. The PEGI is a coregulatory body. So which path should SMCs follow? It is very tempting to use legal concepts and metaphors, but then, we are in the domain of law. Instead of legal concepts and human rights, a separate set of norms and concepts could be created for SMCs.

Sources

Abderrahmane Azzi: "Avoiding Imperialism: Merging the Global and the Local" in Handbook of Global Media Ethics. Edited by Stephen J. A. Ward, Springer, 2021.

Julia Black: "Constitutionalising self-regulation" (1996) 59 Modern Law Review 24.

Chris Frost: Media Ethics and Self-Regulation. London, 2000.

Lorenzo Gradoni: "Twitter Complaint Hotline Operator: Will Twitter Join Meta's Oversight Board?" VerfBlog, 2022/11/10, https://verfassungsblog.de/musk-ob/, DOI: 10.17176/20221110-095537-0.

Christopher T. Marsden: Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace. Cambridge: Cambridge University Press.

Riku Neuvonen: "Internet and Self-Regulation: Media Councils as Models for Social Media Councils?" in: Matthias C. Kettemann (ed.), How Platforms Respond to Human Rights Conflicts Online. Best Practices in Weighing Rights and Obligations in Hybrid Online Orders. Hamburg: Verlag Hans-Bredow-Institut, 2022.

Riku Neuvonen: Sananvapaus, joukkoviestintä ja sääntely. Helsinki, 2005.

Teresa Quintel and Carsten Ullrich: "Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond." in: Bilyana Petkova and Tuomas Ojanen (eds), Fundamental Rights Protection Online the Future Regulation of Intermediaries. Edward Elgar, 2020.

Lennart Weibull and Brit Börjesson: Publicitiska seder. Falun, 1995.

Lindsay Wuller: "Losing the Game." (Winter 2013) 57 Saint Louis University Law Journal, no. 2 457–492.

Social Media Councils in the European Constitutional Framework

Giovanni Di Gregorio
CATÓLICA GLOBAL SCHOOL OF LAW, LISBON, PORTUGAL

Introduction

The introduction of Social Media Councils (SMCs) has captured the attention on a global scale. The attempt of Meta to set up an independent oversight has been one example of the steps towards institutionalising internal systems of review in the field of content moderation. SMCs are increasingly called upon to make decisions on the governance of online content, thus promising to provide answers to the limits of content moderation as a global system.

Still, the introduction of SMCs is not neutral from a constitutional point of view. First, Social Media Councils are boards conceived within the private sector, and this setting primarily influences the scope of constitutional law that traditionally follows a vertical logic in the relationship between freedoms and powers. Second, SMCs adopt and mirror constitutional narratives and procedures, particularly focused on human rights, to make decisions and provide recommendations. This process influences how constitutional values such as freedom of expression or access to justice are interpreted in the digital age not only by law-makers and courts but also by private actors. Third, the scope of SMCs is usually global. Such a perspective leads to a tendency to not consider the constitutional nuances of the local dimension. Despite the universal dimension of human rights, their protection and enforcement are connected to regional and national nuances.

This contribution aims to trigger a conversation about the constitutional role of SMCs. By adopting a European constitutional perspective, this contribution examines the nature of these boards, the use of constitutional narratives and the scope of their activities. The role of SMCs raises peculiar questions in Europe, as underlined by the adoption of the Digital Services Act as the European constitutional reaction to the challenges raised by content moderation.

Private Actors...

SMCs are conceived as private actors. Even if independent from social media, these boards make decisions that are not bound by the rule of law but primarily influenced by the "rule of platforms", such as terms of services, community guidelines and bylaws. Even if SMCs would contribute to increasing awareness of the challenges raised by content moderation while also fixing some pitfalls, they still exercise discretion in making their decisions on human rights.

These actors raise constitutional questions, particularly about how to frame the performance of quasi-public functions by private actors and their private governance on a global scale. Indeed, SMCs contribute to different roles, from adjudication to advisory, but primarily solve conflicts between human rights and private standards. Besides, similarly to constitutional courts, they also make recommendations on community standards in order to align their private standards with human rights. This framework underlines how, in this space, the rule of law has played a limited role so far. The private nature of SMCs narrows the scope of constitutional principles that become only an informal reference to the rule of platforms.

The consolidation of SMCs constitutes another example of the path towards the consolidation of the "rule of platforms". This process is outside the traditional scope of constitutional law. Rights and procedural safeguards directly apply only to public actors. In the lack of any regulation, it is not possible to require private actors to respect constitutional rights such as freedom of expression or, broadly, the rule of law. Constitutions are a critical part of the social contract that, for good reasons, limit governmental powers and ensure individual freedoms from interference by public authorities. Nonetheless, this mission has traditionally focused on limiting the authority of public actors rather than private powers.

At the same time, European constitutionalism is not based on a rigid vertical model. Both the European Convention on Human Rights and the European Charter of Fundamental Rights provide a constitutional system that can react to the challenges raised by private powers. The abuse of rights clause is an example of European constitutionalism that does not tolerate that the protection of economic freedoms turns into justifications to exercise (private) powers. This approach has led to extending constitutional safeguards into horizontal relationships, thus underlining how the European constitutional framework tends to react against unaccountable exercises of powers. Besides, the challenges for fundamental rights driven by a private system of adjudication can also trigger the positive obligation of States to protect human rights, thus leading to a new regulatory landscape for SMCs.

...Watching Over Human Rights

SMCs can play a critical role in fostering oversight, particularly in those areas where the exercise of public authority is absent or excessive, or even when States do not have the power to negotiate safeguards to increase transparency and accountability in content moderation. Nonetheless, these self-regulatory bodies contribute to the institutionalisation of procedural rules that raise questions about access and due process in these private spaces.

In this case, the primary constitutional issue is that human rights work as parameters to review social media decision-making and assess terms of services and community guidelines. This process entails that the subject of (constitutional) scrutiny is not primarily the rule of law but private standards. In this case, SMCs lead to the hybridisation of human rights enforcement and judicial review. Unlike the traditional rule where fundamental rights play the role of parameters to review legislation, in this case, private standards are de facto equivalent to the law, thus confirming the institutionalisation of the "rule of platforms" as the primary system of governance. In other words, these private rulebooks constitute a quasi-legal basis according to which platforms exercise their powers, without relying on public authorities.

Besides, even if SMCs' oversight is based on human rights law, it is not possible to exclude the influence resulting from private standards defined in community guidelines. This is a critical point about participation and representation in the digital environment. If private standards are developed outside a system of transparency and accountability but are only legitimised by users' adherence, the role of oversight, and review, loses its power, and its primary role, i.e. to ensure that constitutional values agreed by a community limit discretionary decisions. If private standards are still defined from the top, any system of oversight would only be an important but empty exercise. This issue also suggests why it is challenging to frame the rise of SMCs as an expression of societal constitutionalism, considering SMCs more as a transnational attempt to institutionalise procedures and rules on a global scale.

The consolidation of SMCs oversight contributes to the marginalisation of public actors in ensuring the enforcement of rights and freedoms online. This framework leads to looking at SMCs as additional

system of adjudication to protect fundamental rights in the digital age. These boards also provide perspectives on the hybridisation of the global approaches to online speech. This additional and alternative system of oversight and adjudication constitutes another call for European constitutionalism that, as already stressed, tends to be intolerant to decision-making processes affecting constitutional values outside the scope of public safeguards. Likewise, the model advanced by SMCs could also be a "positive" trigger for public actors to improve the enforcement and oversight of fundamental rights in the digital age.

Clashing Territorial Dimensions

The predominance of a global approach to SMCs also leads to another constitutional question. SMCs usually tend to operate on a global scale, and this approach is also the reasons why human rights law is particularly appealing for their decision-making. The Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and regional systems such as the European Convention on Human Rights, are only some of the instruments that protect human rights, thus recognising their universal dimension.

Nonetheless, the protection of rights and freedoms is far from being equal geographically. Already in the European context, human rights are subject to different institutional and political dynamics shaping their protection. Even more importantly, human rights do not always overlap with the protection of fundamental rights as recognised by supranational systems and national constitutions. Without addressing the relationship between human rights and fundamental rights, it is critical to stress how the use of a universal approach by SMCs lead to reducing the importance of local constitutional nuances. Taking as example the EU and the US as two consolidated constitutional democracies, the protection of fundamental rights and freedoms do not always overlap and, in some cases, leads to opposite constitutional paths as underlined by the protection of free speech.

These constitutional nuances are underlined by the European path toward mitigating social media powers by proceduralising content moderation. The Digital Services Act is just an example of a path towards the injection of due process safeguards in the activities of online platforms taking decisions on content on a global scale. While the Union is at the forefront of a new constitutional phase addressing the challenges raised by the exercise of private powers in the digital age, the US has not shown the same concern, following an opposite path. For instance, the Communication Decency Act immunises online intermediaries, including modern online platforms, from liability when moderating users' content.

Even if SMCs take into account global community standards and human rights, they address local challenges as underlined by the cases addressed by the Meta Oversight Board. In this case, in order to represent constitutional nuances, the proliferation of regional SMCs can take into account the protection of fundamental rights into the local dimension. This approach would also lead to overcome the limited composition of these bodies that usually are not adequate to represent the nuances of rights and freedoms on a global scale. Besides, this change would also lead to increasing impact. The Meta Oversight Board is still a small move in the field considering the very few cases that this board will likely consider and also that the board provides an ex-post remedy which is not able to address the large amounts of cases raised by users.

Constitutional Perspectives

SMCs will not probably solve the primary challenges of content moderation such as content monetisation and profiling, or the biases of artificial intelligence moderating content. Nonetheless, these

bodies can play a critical role in improving the process of content moderation, particularly making the governance of online speech more transparent and accountable.

The question is how and to what extent SMCs will be part of the new strategy of the Union to proceduralise content moderation. The introduction of the Digital Services Act has defined a landmark step by providing procedural safeguards and redress mechanisms, and it is likely to play an important role also in relation to SMCs. For instance, the obligation for online platforms to take into account the protection of fundamental rights in their terms of services would also impact how SMCs review the compatibility of private standards with human rights. This rule will not only contribute to making the rule of law part of the discussion but also to increase the nuances of SMCs decision-making.

The European approach contributes to the expansion of SMCs. Even if social media are not required to establish SMCs, still these bodies can show accountability while assessing and mitigating risks such as the spread of harmful content. In this case, in Europe, SMCs are likely to become critical parts of social media architecture, that, from a constitutional perspective, is also governed by the rule of law.