**Selected Papers of #AoIR2022:**
**The 23rd Annual Conference of the**
**Association of Internet Researchers**
Dublin, Ireland / 2-5 Nov 2022

# POLICING PLATFORMS: ADDRESSING POWER AND INEQUALITIES IN PLATFORM POLICIES

Christian Katzenbach
ZeMKI Centre for Media, Communication and Information Research, University of Bremen

Dennis Redeker
ZeMKI Centre for Media, Communication and Information Research, University of Bremen

João Carlos Magalhães
University of Groningen

Adrian Kopps
ZeMKI Centre for Media, Communication and Information Research, University of Bremen and Alexander von Humboldt Institute of Internet and Society, Berlin

Tom Sühr
Technical University of Berlin and Harvard Business School

Robyn Caplan
Data & Society, New York

Paloma Viejo Otero
Dublin City University

Edoardo Celeste
Dublin City University

Nicola Palladino
Trinity College Dublin

Kinfe Yilma
Addis Ababa University

Suggested Citation (APA): Katzenbach, C., Redeker, D., Magalhães, J. C., Kopps, A., Sühr, T., Caplan, R., Viejo Otero, P., Celeste, E., Palladino, N., & Yilma, K. (2022, November). *Policing Platforms: Addressing Power and Inequalities in Platform Policies.* Panel presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

Social media platforms are ubiquitous in today's Internet. Accepting that these digital intermediaries "govern the Internet" (Suzor 2019: 168), it matters which rules they are bound by and bind themselves to. Platforms had for a long time successfully positioned themselves in the "sweet spot" between beneficial legislative protections and a remarkable absence of obligations (Gillespie 2010: 348), yielding little need to take direct responsibility for the content of users. During the last decade however, and specifically since 2016, public and policy pressure has pushed platforms to become something different: not the allegedly neutral tech companies, but powerful intermediaries responsible for the functioning of public discourse and democracy. As a result, platforms have developed (more or less) elaborate policies that in effect govern *who can say what when and where* on the Internet. Platforms have struggled to develop their positions and processes for handling contested and delicate issues such as hate speech and misinformation, and the policies are still changing regularly. This highly dynamic situation indicates a low level of stability and institutionalization (Barret & Kreiss 2019; Katzenbach 2021; Katzenbach et al. 2021). Taken together, this makes platform policies a key site to investigate today's internet power relations, and the tech giants' strategies to "re-fashion the world in their own image" (cf. CfP). It is the locus where the power of the "new governors" (Klonick 2019) becomes manifest, and yet it is still subject of negotiation and adoption.

This panel examines how platform policies have come to be the way they are today, the influence of legal principles and political processes on them, and their enactment in practices and organizational processes. By studying the written (and unwritten) policies of social media platforms across different topical areas such as copyright, hate speech and account verification and on different levels of content moderation, we aim to better understand how these intermediaries indeed govern the Internet. The (1) first paper on *copyright content moderation* investigates platform policies with regard to copyright in detail across fifteen platforms and a time period of more than ten years. The analysis shows that platform regulation in this area has evolved drastically over this time period with changing normative types (rights, obligations, expectations, principles) and subjects of moderation, resulting in a complexification and commodification of copyright content moderation. The (2) second paper provides a *typology of verification policies* across several major platform companies, building on the examination of how platforms as diverse a eBay, Pornhub, Twitter, Airbnb and others, have developed processes and policies to classify some users, things, and places as 'official,' or 'authentic.' The analysis shows how verification is not just a matter of identity authentication, but that verification policies also importantly signal organizational and institutional relationships between platforms and their user groups that can confer significant material and social benefits. The (3) third develops a *critical approach to the relationship between human reviewers and AI* in content moderation adopting a decolonial perspective. By looking at the entanglements in the enforcement of platform policies, this paper explores the actual means by which human moderators input AI systems with their decisions and asks to what extent this relationship is one of AI-employee-collaboration, a form of colonization of the imagination, or of knowledge expropriation. The (4) fourth paper

adds a *socio-legal perspective to the debate on the influences on (and principles embedded in) platform content moderation policies*. Using the theoretical framework of digital constitutionalism, the paper examines to what extent either human rights standards or demands by civil society activism can represent a fitting standard for platform content moderation policies. Analyzing the policies and moderation practices of YouTube, Twitter, TikTok and Meta, the paper finds a great degree of institutional isomorphism between them but also significant discrepancies with regard to the adoption of these standards. The paper also suggests that new actors like Meta's Oversight Board already play an important role in the translation process.

Taken together, the papers of this panel analyze crucial power dynamics and inequalities embedded within and extending beyond (social media) platform policies. For this, the panel brings together different disciplines and methods and fosters a productive multidisciplinary conversation and methodological exchange. This examination of the social, political, legal and economic underpinnings of recent changes in platform policies from a global perspective will allow us to better understand the ability of platforms to "re-fashion the world in their image" and to foster change.

**References**

Barrett, B., & Kreiss, D. (2019). Platform transience: Changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, 8(4). https://policyreview.info/articles/analysis/platform-transience-changes-facebooks-policies-procedures-and-affordances-global

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, 12(3), 347-364.

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society*, 8(2). https://doi.org/10.1177/20539517211046182

Katzenbach, C., Magalhães, J. C., Kopps, A., Sühr, T. & Wunderlich, L. (2021). *The Platform Governance Archive*. Alexander von Humboldt Institute for Internet and Society. https://doi.org/10.17605/OSF.IO/XSBPT

Klonick, K. (2019). The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, 129, 2418.

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge: Cambridge University Press.

*Paper 1*

# COMPLEXIFICATION AND CONCENTRATION IN PLATFORM POWER: CHANGES IN COPYRIGHT CONTENT MODERATION AND POLICIES ACROSS 10 YEARS AND FIFTEEN PLATFORMS

Christian Katzenbach
ZeMKI Centre for Media, Communication and Information Research, University of Bremen

João Carlos Magalhães
University of Groningen

Adrian Kopps
ZeMKI Centre for Media, Communication and Information Research, University of Bremen and Alexander von Humboldt Institute of Internet and Society, Berlin

Tom Sühr
Technical University of Berlin and Harvard Business School

## Introduction

Platform governance and content moderation are a key site of the struggle to define the Internet, its power relations, and the tech giants' strategies to "re-fashion the world in their own image" (cf. CfP). While the politicization of these questions with the advent of the tech lash 2015-17 and its focus on hate speech and misinformation might be considered a rather recent phenomenon, controversies on the responsibility and liability of platforms and service providers have been around since the commercialization of the Internet in the 1990s. Copyright has been a key topic among these long-standing debates, and continues to provide a rich source of contestation. Massive public protests have, for example, accompanied the adoption of the recent copyright directive of the European Union, with critics anticipating forced censorship by the tech platforms and the end of the Internet as we know it, once the directive comes into effect. That is why it constitutes a particular instructive topic to study from a long-term perspective how power relations shape the Internet as we know it.

## Investigating platform power by looking at copyright content moderation

While there is now ample research on different aspects of platform governance, there is surprisingly little empirical research on copyright content moderation. Acknowledging that platform policies, or “platform law” (Kaye 2019), are where the ways in which “digital intermediaries govern the Internet” (Suzor 2019: 168) become manifest, researchers have, for example, studied policies on hate speech (Siapera & Viejo Otero, 2021), misinformation (Keulenaar et al., 2021), or child abuse (Tarvin & Stanfill, 2022). Others have systematically compared policies across platforms and topics (Buckley &

Schaefer, 2021) and highlighted the dynamic nature of policies and their responsivity to public and policy pressure (Barrett & Kreiss, 2019; Katzenbach, 2021). With regard to copyright, platform policies are not yet well researched. Given that earlier work has identified structural overblocking of content (ie. blocking and deletion of content legally does not constitute copyright infringement) by platforms and providers (Urban et al., 2018; Erickson & Kretschmer, 2018; Lester & Pachamanova, 2017; Bar-Ziv & Elkin-Koren, 2018), there is evident need to more closely examine the copyright policies of platforms and their changes over time. Copyright content moderation is potentially one of the key mechanisms to execute structural power over the circulation of content and creativity online. Powerful actors such as media companies have in the past repeatedly mobilized copyright claims and enforcement mechanisms to strengthen their positions and to disadvantage independent creators and publishers as well as users and citizens.

Against this background, this paper describes a longitudinal study on how platforms have handled copyright-related questions in their content policies, that is the Terms of Services, Community Guidelines, and Copyright Policies. We study these questions across a diverse selection of social media platforms and over a time period of more than ten years, in order to understand the dynamics and changes across time and across different platforms. In the analysis we (1) explore which kinds of public documents and rules fifteen major and niche platforms have adopted to regulate copyright content moderation, we then (2) examine in detail how the rules of a selection of six of these platforms changed over time, and then (3) finally discuss and compare the enforcement of these policies in platforms' automated copyright content moderation systems.

**Results: Complexification and concentration of platform power**

Our analysis shows that copyright content moderation has evolved drastically over the recent two decades. These results specifically show along our two dimensions of analysis: the different *normative types* platforms use to regulate, moderate and evaluate content, including: rights, obligations, expectations, principles, and procedures; and the differing *subjects* of copyright content regulation, including: infringement avoidance, manual content removal, automated moderation, disputes, penalties, exceptions, transparency, and monetisation.

In sum, the results suggest that two dual processes seem to mark the development of platforms' copyright content moderation over time. Firstly, there has been a process of complexification and opacification. Our empirical work indicates that virtually all 15 platforms' T&Cs have become more intricate, in various ways and to different extents, a process that was deepened by the emergence of automated copyright content moderation systems. It seems evident that the structures we studied became increasingly harder to understand and, sometimes, even to observe. Such opacification was neither necessary nor necessarily justifiable, the paper argues. We characterize the second process as platformisation and concentration. Our research demonstrates that platforms have altered their rules so as to subsume copyright content moderation to their own interests, logics and technologies. As with complexification, platformisation

has been experienced differently by different platforms and deepened by the rise of automated systems. This has seemingly led to a concentration of power in the hands of not only platforms themselves but large rights holders as well, the paper concludes.

**References**


Barrett, B., & Kreiss, D. (2019). Platform transience: Changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, 8(4). https://policyreview.info/articles/analysis/platform-transience-changes-facebooks-policies-procedures-and-affordances-global.

Bar-Ziv, S., & Elkin-Koren, N. (2018). Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown. *Connecticut Law Review* 50.

Buckley, N., & Schafer, J. S. (2022). 'Censorship-free' platforms: Evaluating content moderation policies and practices of alternative social media. *For(e)Dialogue*, 4(1). https://doi.org/10.21428/e3990ae6.483f18da.

Erickson, K., & Kretschmer, M. (2018). 'This Video is Unavailable': Analyzing Copyright Takedown of User-Generated Content on YouTube. *Journal of Intellectual Property, Information Technology and E- Commerce Law (JIPITEC)*, 9(1), 75–89.

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society*, 8(2). https://doi.org/10.1177/20539517211046182.

Kaye, D. (2019). Speech Police – The Global Struggle to Govern the Internet. New York: Columbia Global Reports.

Keulenaar, E. de, Burton, A. G., & Kisjes, I. (2021). Deplatforming, demotion and folk theories of Big Tech persecution. *Fronteiras - Estudos Midiáticos*, 23(2), 118–139. https://doi.org/10.4013/fem.2021.232.09

Lester, T., & Pachamanova, D. (2017). The Dilemma of False Positives: Making Content ID Algorithms more Conducive to Fostering Innovative Fair Use in Music Creation. *Entertainment Law Review*, 24.

Siapera, E., & Viejo-Otero, P. (2021). Governing Hate: Facebook and Digital Racism. *Television & New Media*. https://doi.org/10.1177/1527476420982232.

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge: Cambridge University Press.

Tarvin, E., & Stanfill, M. (2022). "YouTube's predator problem": Platform moderation as governance-washing, and user resistance. *Convergence*, 13548565211066490. https://doi.org/10.1177/13548565211066490.

Urban, J., Karaganis, J., & Schofield, B. (2018). Takedown in Two Worlds: An Empirical Analysis. *UC Berkeley Public Law Research Paper* No. 2755628. https://doi.org/10.31235/osf.io/mduyn.

*Paper 2*

## THE BUSINESS OF TRUST: A TYPOLOGY OF VERIFICATION IN PLATFORM GOVERNANCE

Robyn Caplan
Data & Society, New York

As platforms come to embrace their role as *mediators* of the Internet, they are increasingly using tools like verification (i.e. the blue checkmark used by Twitter and Instagram) as a way to distinguish between official and unofficial sources. For instance, the online adult video platform, Pornhub, announced they had removed all unverified videos, limiting uploads to verified users only (Cole, 2020). The move followed an investigative opinion piece by *The New York Times*'s Nicholas Kristof that followed the lives of sexual abuse victims whose videos were uploaded to the site. Kristof alleged that rape videos, including child rape videos, were allowed to remain and spread on the site unchecked (Kristof, 2020). In response, major payment companies like Mastercard and Visa began their own investigations, eventually announcing they would stop processing payments with Pornhub (Robertson, 2020). Pornhub's move to "verified users only" means that uploads can only come from official content partners and members of their "Model Program" (Pornhub, 2020). Pornhub Verified, which requires content partners to send in a current photo ID for performers, has drastically reduced the amount of content available on the site. All the same, victims of child sex trafficking who had their videos uploaded to the site have argued in a recent lawsuit that the new rules have not gone far enough, noting that the company has made "no effort" to verify the identity of all performers in a video (Ross, 2021).

Pornhub, however, is not alone in this move towards prioritizing verified users and content as a way to mitigate content concerns. Platforms have begun embracing more publicly their role as *mediators* of information, and between interest groups vying for status online. What is happening on Pornhub and many other platforms is part of this broader shift: many, even most, platforms are using "verification" as a way to distinguish between sources, often framing these efforts within concerns about safety and trustworthiness. For instance, Airbnb announced in 2019 that it would verify all of its listings (Yaffe-Bellany, 2019), including the accuracy of photographs, addresses, and other information posted by hosts about themselves and their properties. Tinder has rolled out a blue checkmark verification system to deter catfishing, asking users to take selfies in real time and match poses in sample images (Carman, 2020). Perhaps in recognition of the importance verification will play in the future of the internet, Twitter

has opened a draft of their new verification system to public comment (Twitter, n.d.; Twitter Inc., 2020). As work by Caplan & Gillespie (2020) has demonstrated, other platforms where legacy media and amateur content creators converge, such as YouTube, have different content moderation rules and processes for different user groups.

## A Typology of Verification Practices and Policies

This paper explores how platform companies use verification policies as a way to differentiate between users and goods over their networks. Verification will play an important role in the future of platform governance, and how platforms implement verification policies and processes will have important consequences for participation, inclusion, and diversity online. This paper is an overview of verification policies and practices across major platform companies. This includes social media platforms such as Weibo, Instagram, Twitter, and Tiktok using blue checkmarks to verify identities, organizations, or signal official sources, e-commerce platforms such as eBay and Amazon verifying sellers and products, transportation platforms like Uber who use tools like facial recognition to verify the identity of drivers, and platforms like Airbnb, that are using verification as a way to review the "quality" of goods for consumers. This paper uses publicly available documents from platform companies, including community guidelines, user interfaces for verification, terms of service agreements, and posts from corporate blogs and websites, as well as other public statements made by company representatives. I use the WayBackMachine to understand how these policies have changed over time. I also rely on search engine and social media trade reporting.

This paper finds a broad range of verification policies and processes across these major platforms, and holds that verification is not just a matter of identity authentication (van der Nagel, 2020). Rather, verification policies signal organizational and institutional relationships between platforms and their user groups that can confer significant material and social benefits. Examining verification policies in detail, and doing so through this comparative approach, provides a lens into understanding a platform's economic and institutional ties as they mediate between the users and organizations on their networks.

## References

Carman, A. (2020, January 23). *Tinder will give you a verified blue check mark if you pass catfishing test.* Retrieved from The Verge: https://www.theverge.com/2020/1/23/21077423/tinder-photo-verification-blue-checkmark-safety-center-launch-noonlight

Cole, S. (2020, December 14). *Pornhub just purged all unverified content from the platform.* Retrieved from Vice: https://www.vice.com/en/article/jgqjjy/pornhub-suspended-all-unverified-videos-conteng

Kristof, N. (2020, December 4). *The Children of Pornhub.* Retrieved from The New York Times:
https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html

Pornhub. (2020, December). *The latest on our commitment to trust and safety.* Retrieved from Pornhub: https://www.pornhub.com/blog/11422

van der Nagel, E. (2020). Embodied verification: Linking identities and bodies on NSFW Reddit. In K. Warfield, C. Abidin, & C. Cambre, *Mediated Interfaces: The Body on Social Media.* New York: Bloomsbury.

Ross, S. (2021, February 13). *New lawsuit against Pornhub alleges improvements don't go far enough.* Retrieved from CTV News.
https://montreal.ctvnews.ca/new-lawsuit-against-pornhub-alleges-improvements-to-the-site-don-t-go-far-enough-1.5308001

Twitter. (n.d.). *About verified accounts.* Retrieved from Twitter Help Center:
https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Twitter Inc. (2020, November 24). *Help us shape our new approach to verification.* Retrieved from Twitter Blog:
https://blog.twitter.com/en_us/topics/company/2020/help-us-shape-our-new-approach-to-verification.html

Yaffe-Bellany, D. (2019, November 6). *Airbnb to verify all listings, C.E.O. Chesky says.* Retrieved from The New York Times:
https://www.nytimes.com/2019/11/06/business/airbnb-verify-listings.html

*Paper 3*

## HUMAN REVIEWERS & AUTOMATIC DETECTION SYSTEMS: ARE THEY AI-EMPLOYEE COLLABORATION, COLONIZATION OF THE IMAGINATION OR KNOWLEDGE EXPROPRIATION?

Paloma Viejo Otero
Dublin City University

The paper develops a critical approach to the relationship between Human Reviewers and AI in content moderation adopting a decolonial perspective. Specifically, we ask to what extent the relationship between Human Reviewers and AI systems is one of employee collaboration, a form of colonization of the imagination, or a form of knowledge expropriation. To formulate this thought, the paper draws upon the

influences from Anibal Quijano's theory of the coloniality of power and the question of knowledge as a product of a subject-object relation (2007). It also draws on Eugenia Siapera's work in AI Content Moderation, Racism and (de) Coloniality (2021), and Sarah Roberts work on Human Review Moderators (2019).

Content moderation is an essential and structural feature of platforms that responds to a security rationale by which large volumes of harmful content should be efficiently removed (Siapera 2021, Siapera and Viejo Otero 2021). In relation to content moderation, Mark Zuckerberg announced in 2017 the hiring of ten thousand employees, 'for safety and security, with the possibility of extending that number to twenty thousand'. In January 2018, Mark Zuckerberg reiterated that 'to prevent hate speech and ensure the security of the platform, Facebook invests in staff and technology in equal parts so that Facebook invests around 14,000 people working across communities'. By 2019, Facebook had a total of '30,000 people working at Facebook just for safety and security where half of those are content reviewers'. Therefore, it could be argued that Mark Zuckerberg's actions indicate that Human Reviewers and AI's relationship is indissoluble in the eyes of tech companies, and that content moderation is a growing labour market for individuals who possess specific technical, linguistic and cultural knowledge.

Indeed, Social Media platforms are interested in presenting Human Reviewers and Automated systems under the lens of mutual collaboration. Human reviewers are guided by content standards and their task is to review all posts and publications flagged as potentially harmful. These teams are often determined by the languages and cultures of the individuals in these teams, so that moderators are people with necessary cultural and linguistic knowledge to decide over content. Whereas the role of automatic systems is to faster eliminate material that repeatedly appears on the platform and that platforms recognise as harmful. Considering that human moderation was implemented before Automatic Systems, it could be suggested that human moderators have continuously generated knowledge that the companies have accumulated to input their algorithms.

Different, however, is how human reviewers experiment their working relationship with AI. Often, literature refers to the relationship between Humans and AI systems as 'human-AI partnership', 'human-AI teaming' (Nguyen et al 2018, Barro & Davenport 2019) or human-AI collaboration (Ashktorab 2020). This literature assumes that humans cooperate and align their interest with the interest of Social Media Platforms around moderation and security (Chandrasekharan et al 2017, Lai et al 2022). However, there is a dearth of empirical knowledge about if human teams perceive this relationship as a collaboration, as knowledge expropriation (Heiman and Nickerson 2004, Quijano 2007) or even as a colonization of the imagination (Quijano 2007).

In light of this gap, the present paper questions if AI's relationship in Content Moderation can be leveled as 'human-AI teaming', which implies equal collaboration, or as 'human-in-the-loop' by which AI receives input from the human to improve its performance without human acknowledgement and consent. To answer this question, this paper empirically relies on a series of in- depth interviews with human moderators. Questions explore human reviewer's daily activity, how and if they actively input AI

systems, and if so, what are the actual means by which this occurs. Data from the human reviewers will be analysed through the analytic lenses of team-AI collaboration, expropriation of knowledge and colonisation of the imagination. In doing so, this paper aims to contribute to deciphering the extent by which colonial practices operate in the social media environment.

**References**


Ashktorab, Z., Liao Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW2. 1–20.

Barro, S., & Davenport, T. H. (2019). People and machines: Partners in innovation. *MIT Sloan Management Review*, *60*(4), 22-28.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., , Eisenstein, J., & Gilbert, E. (2017). You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of ACM Human Computer Interaction 1, Computer-Supported Cooperative Work and Social Computing*, Article 31.

Chowdhury, S., Budhwar, P., Dey, P. K., Joel-Edgar, S., & Abadie, A. (2022). AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research*, *144*, 31-49.

Heiman, B. A., & Nickerson, J. A. (2004). Empirical evidence regarding the tension between knowledge sharing and knowledge expropriation in collaborations. *Managerial and Decision Economics*, *25*(6-7), 401-420.

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation.

Nguyen, An T., Kharosekar, A., Krishnan. S., Krishnan,S., Tate, E., Wallace, B.. & Lease M. (2018). Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 189–199.

Quijano, A. (2007). Coloniality and modernity/rationality. *Cultural Studies*, *21*(2-3), 168-178.

Roberts, S. T. (2019). *Behind the Screen*. New Haven: Yale University Press.
Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5), 1-35.

Siapera, E. (2021). AI Content Moderation, Racism and (de) Coloniality. *International Journal of Bullying Prevention*, 1-11.

Viejo-Otero, P. (2022). *Governing Hate: Facebook and Hate Speech*. PhD Thesis, Dublin City University.

*Paper 4*

# DIGITAL CONSTITUTIONALISM AND PLATFORM POLICIES: TOWARDS A GLOBAL STANDARD?

Edoardo Celeste
Dublin City University

Nicola Palladino
Trinity College Dublin

Dennis Redeker
ZeMKI Centre for Media, Communication and Information Research, University of Bremen

Kinfe Yilma
Addis Ababa University

## A Dilemma for Social Media Content Governance

One of the main issues of global content governance on social media relates to the definition of the rules governing online content moderation worldwide (Gillepsie 2018). One could think that it would be sufficient for online platforms to refer to existing international human rights standards. However, a more careful analysis shows not only that international law provides exclusively general principles, which do not specifically address the context of online content moderation. But also that a single human rights standard does not exist as even identical provisions and principles are interpreted by courts in different ways across the world. This is one of the reasons why, since their birth, major social media platforms have set their own rules, adopting their own peculiar language, values and parameters in their platform policies (Suzor 2019). Yet, at the same time, this normative autonomy too has raised serious concerns. Why should private companies establish the rules governing free speech online? Is it legitimate to depart from minimal human rights standards and impose more (or less) stringent rules?

The current situation exposes a dilemma for online content governance that seriously affects the operations of social media companies and impacts on the exercise of fundamental rights by users as well as digital policy strategies. On the one hand, if social media platforms simply adopted international law standards, they would be

compelled to operate a choice on which model to follow – for example, between a US freedom of expression-dominated approach or a European-style standard, which balances freedom of expression with other social values. Moreover, they would also need to put in place a mechanism able to translate, or 'operationalise', such general standards in their platform policies (Suzor 2018). On the other hand, where social media platforms adopt their own values, rules and terminology in their policy documents, thus departing from international law standards, they are accused of censorship or laxity, intrusiveness or negligence.

**Translating International Human Rights Standards into Platform Policies?**

In this paper we address this conundrum. The paper investigates this topic from a multidisciplinary perspective, drawing from the expertise of the authors in law, political science and communication studies. We argue that the key to solving this dilemma lies in the capacity to define specific principles and values for as the foundation of platform policies, a task which is part of the broader process of constitutionalising the digital society and is the main aim of the new ideological movement called 'digital constitutionalism' (Celeste 2019). To this end, the paper will explore civil society declarations articulating constitutional principles related to content governance, the so-called Internet bills of rights (Celeste 2018).

We argue that the potential of international human rights law in offering much needed normative guidance to content governance is circumscribed by three interrelated factors. First, international human rights law is – by design – State-centered and hence does not go a long way in attending to human rights concerns in the private sector (Lwin 2020). Second, international human rights law standards are couched in general principles, and hence, less suited to apply in the context of platform content moderation which requires a rather granular and dynamic system of norms. Third, the generic international content governance standards have not adequately been unpacked by relevant adjudicative bodies to make them fit for purpose to the present realities of content moderation and for inclusion in platform people in particular.

Civil Society declarations constitute a source of normative standards for platform policies that have been so far neglected by the scholarship. Over the past few years, a series of initiatives have emerged at societal level, and especially among civil society groups, to articulate rights and principles for the digital age (Redeker et al. 2019). The output of these efforts mostly consists of non-legally binding declarations, often intentionally adopting a constitutional tone and therefore termed 'Internet bills of rights'. They can be considered as expressing the "voice" of communities that struggle to propose an innovative message within traditional institutional channels: one of the layers of the complex process of constitutionalisation (Celeste 2021) that is pushing towards reconceptualising core constitutional principles in light of the challenges of the digital society. Moreover these texts can provide an idea of which human rights standards are currently promoted by the communities within which social media platforms operate. This section illustrates the findings of a content analysis on a

database of 40 Internet bills of rights including principles related to online content governance.

Finally, we illustrate how social media platforms deal with the content governance dilemma outlined previously. Both international human rights law and civil society documents speak to platform policies and those who design them. By comparing four major social media platforms - Meta, Twitter, TikTok and YouTube - in order we examine how substantive and procedural principles entailed in their policy documents relate to the standards stemming from international human rights law and internet bills of rights. This analysis is based on primary, publicly available data provided by the platforms, including their policy documents and enforcement reports, and secondary literature. This section examines differences and similarities between the Twitter Rules, Meta's Community Standards, TikTok Community Guidelines and YouTube Community Guidelines. A comparison of principles found in civil society declarations and international human rights law with the platform policies shows great divergence of adoption of the former. The analysis also shows how emerging actors, specifically Meta's Oversight Board, engage in the translation work of external standards into platform policies.

**References**

Celeste, E. (2018). Terms of service and bills of rights: New mechanisms of constitutionalisation in the social media environment? *International Review of Law, Computers & Technology, 33*(2), 122-138.

Celeste, E. (2019). Digital constitutionalism: A new systematic theorisation. *International Review of Law, Computers & Technology, 33*(1), 76-99.

Celeste, E. (2021). The constitutionalisation of the digital ecosystem: Lessons from international law. *Max Planck Institute for Comparative Public Law & International Law (MPIL) Research Paper*, (2021-16).

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

Lwin, M. (2020). Applying International Human Rights Law for Use by Facebook. *JREG Bulletin*, 38, 53.

Redeker, D., Gill, L., & Gasser, U. (2018). Towards digital constitutionalism? Mapping attempts to craft an Internet bill of rights. *International Communication Gazette, 80*(4), 302-319.

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society, 4*(3), 1-11.

Suzor, N. (2019). *Lawless: The Secret Rules that Govern our Digital Lives*. Cambridge: Cambridge University Press.